# Chapter 3
# Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-$d'$, Response-Specific Meta-$d'$, and the Unequal Variance SDT Model

**Brian Maniscalco and Hakwan Lau**

**Abstract** Previously we have proposed a signal detection theory (SDT) methodology for measuring metacognitive sensitivity (Maniscalco and Lau, Conscious Cogn 21:422–430, 2012). Our SDT measure, meta-$d'$, provides a response-bias free measure of how well confidence ratings track task accuracy. Here we provide an overview of standard SDT and an extended formal treatment of meta-$d'$. However, whereas meta-$d'$ characterizes an observer's sensitivity in tracking overall accuracy, it may sometimes be of interest to assess metacognition for a particular kind of behavioral response. For instance, in a perceptual detection task, we may wish to characterize metacognition separately for reports of stimulus presence and absence. Here we discuss the methodology for computing such a "response-specific" meta-$d'$ and provide corresponding Matlab code. This approach potentially offers an alternative explanation for data that are typically taken to support the unequal variance SDT (UV-SDT) model. We demonstrate that simulated data generated from UV-SDT can be well fit by an equal variance SDT model positing different metacognitive ability for each kind of behavioral response, and likewise that data generated by the latter model can be captured by UV-SDT. This ambiguity entails that caution is needed in interpreting the processes underlying relative operating characteristic (ROC) curve properties. Type 1 ROC curves generated by combining type 1 and type 2 judgments, traditionally interpreted in

B. Maniscalco (✉)
National Institute of Neurological Disorders and Stroke, National Institutes of Health, 10 Center Drive, Building 10, Room B1D728, MSC 1065, Bethesda, MD 20892-1065, USA
e-mail: bmaniscalco@gmail.com

B. Maniscalco · H. Lau
Department of Psychology, Columbia University, 406 Schermerhorn Hall, 1190 Amsterdam Avenue MC 5501, New York, NY 10027, USA
e-mail: hakwan@gmail.com

H. Lau
Department of Psychology, UCLA, 1285 Franz Hall, Box 951563 Los Angeles, CA 90095-1563, USA

S. M. Fleming and C. D. Frith (eds.), *The Cognitive Neuroscience of Metacognition*,
DOI: 10.1007/978-3-642-45190-4_3, 

terms of low-level processes (UV), can potentially be interpreted in terms of high-level processes instead (response-specific metacognition). Similarly, differences in area under response-specific type 2 ROC curves may reflect the influence of low-level processes (UV) rather than high-level metacognitive processes.

## 3.1 Introduction

Signal detection theory (SDT; [10, 12]) has provided a simple yet powerful methodology for distinguishing between *sensitivity* (an observer's ability to discriminate stimuli) and *response bias* (an observer's standards for producing different behavioral responses) in stimulus discrimination tasks. In tasks where an observer rates his confidence that his stimulus classification was correct, it may also be of interest to characterize how well the observer performs in placing these confidence ratings. For convenience, we can refer to the task of classifying stimuli as the type 1 task, and the task of rating confidence in classification accuracy as the type 2 task [2]. As with the type 1 task, SDT treatments of the type 2 task are concerned with independently characterizing an observer's type 2 sensitivity (how well confidence ratings discriminate between an observer's own correct and incorrect stimulus classifications) and type 2 response bias (the observer's standards for reporting different levels of confidence).

Traditional analyses of type 2 performance investigate how well confidence ratings discriminate between all correct trials versus all incorrect trials. In addition to characterizing an observer's overall type 2 performance in this way, it may also be of interest to characterize how well confidence ratings discriminate between correct and incorrect trials corresponding to a particular kind of type 1 response. For instance, in a visual detection task, the observer may classify the stimulus as "signal present" or "signal absent." An overall type 2 analysis would investigate how well confidence ratings discriminate between correct and incorrect trials, regardless of whether those trials corresponded to classifications of "signal present" or "signal absent." However, it is possible that perceptual and/or metacognitive processing qualitatively differs for "signal present" and "signal absent" trials. In light of this possibility, we may be interested to know how well confidence characterizes correct and incorrect trials *only* for "signal present" responses, or *only* for "signal absent" responses (e.g. [11]). Other factors, such as experimental manipulations that target one response type or another (e.g. [7]) may also provide impetus for such an analysis. We will refer to the analysis of type 2 performance for correct and incorrect trials corresponding to a particular type 1 response as the analysis of *response-specific*[1] type 2 performance.

[1] We have previously used the phrase "response-conditional" rather than "response-specific" [13]. However, [2] used the terms "stimulus-conditional" and "response-conditional" to refer to

In this article, we present an overview of the SDT analysis of type 1 and type 2 performance and introduce a new SDT-based methodology for analyzing response-specific type 2 performance, building on a previously introduced method for analyzing overall type 2 performance [13]. We first provide a brief overview of type 1 SDT. We then demonstrate how the analysis of type 1 data can be extended to the type 2 task, with a discussion of how our approach compares to that of Galvin et al. [9]. We provide a more comprehensive methodological treatment of our SDT measure of type 2 sensitivity, meta-$d'$ [13], than has previously been published. With this foundation in place, we show how the analysis can be extended to characterize response-specific type 2 performance.

After discussing these methodological points, we provide a cautionary note on the interpretation of type 1 and type 2 relative operating characteristic (ROC) curves. We demonstrate that differences in type 2 performance for different response types can generate patterns of data that have typically been taken to support the unequal variance SDT (UV-SDT) model. Likewise, we show that the UV-SDT model can generate patterns of data that have been taken to reflect processes of a metacognitive origin. We provide a theoretical rationale for this in terms of the mathematical relationship between type 2 ROC curves and type 1 ROC curves constructed from confidence ratings, and discuss possible solutions for these difficulties in inferring psychological processes from patterns in the type 1 and type 2 ROC curves.

## 3.2 The SDT Model and Type 1 and Type 2 ROC Curves

### *3.2.1 Type 1 SDT*

Suppose an observer is performing a task in which one of two possible stimulus classes ($S1$ or $S2$)[2] is presented on each trial, and that following each stimulus presentation, the observer must classify that stimulus as "$S1$" or "$S2$."[3] We may define four possible outcomes for each trial depending on the stimulus and the observer's response: hits, misses, false alarms, and correct rejections (Table 3.1).

(Footnote 1 continued)
the type 1 and type 2 tasks. Thus, to avoid confusion, we now use "response-specific" to refer to type 2 performance for a given response type. We will use the analogous phrase "stimulus-specific" to refer to type 2 performance for correct and incorrect trials corresponding to a particular stimulus.

[2] Traditionally, $S1$ is taken to be the "signal absent" stimulus and $S2$ the "signal present" stimulus. Here we follow [12] in using the more neutral terms $S1$ and $S2$ for the sake of generality.

[3] We will adopt the convention of placing "$S1$" and "$S2$" in quotation marks whenever they denote an observer's classification of a stimulus, and omitting quotation marks when these denote the objective stimulus identity.

**Table 3.1** Possible outcomes for the type 1 task

| Stimulus | Response | |
|---|---|---|
| | "*S*1" | "*S*2" |
| *S*1 | Correct rejection (CR) | False alarm (FA) |
| *S*2 | Miss | Hit |

When an *S*2 stimulus is shown, the observer's response can be either a hit (a correct classification as "*S*2") or a miss (an incorrect classification as "*S*1"). Similarly, when *S*1 is shown, the observer's response can be either a correct rejection (correct classification as "*S*1") or a false alarm (incorrect classification as "*S*2").[4]

A summary of the observer's performance is provided by hit rate and false alarm rate[5]:

$$\text{Hit Rate} = \text{HR} = p(\text{resp} = \text{"}S2\text{"} \mid \text{stim} = S2) = \frac{n(\text{resp} = \text{"}S2\text{"}, \text{stim} = S2)}{n(\text{stim} = S2)}$$

$$\text{False Alarm Rate} = \text{FAR} = p(\text{resp} = \text{"}S2\text{"} \mid \text{stim} = S1) = \frac{n(\text{resp} = \text{"}S2\text{"}, \text{stim} = S1)}{n(\text{stim} = S1)}$$

where $n(C)$ denotes a count of the total number of trials satisfying the condition $C$.

ROC curves define how changes in hit rate and false alarm rate are related. For instance, an observer may become more reluctant to produce "*S*2" responses if he is informed that *S*2 stimuli will rarely be presented, or if he is instructed that incorrect "*S*2" responses will be penalized more heavily than incorrect "*S*1" responses (e.g. [12, 22]); such manipulations would tend to lower the observer's probability of responding "*S*2," and thus reduce false alarm rate and hit rate. By producing multiple such manipulations that alter the observer's propensity to respond "*S*2," multiple (FAR, HR) pairs can be collected and used to construct the ROC curve, which plots hit rate against false alarm rate (Fig. 3.1b[6]).

On the presumption that such manipulations affect only the observer's *standards* for responding "*S*2," and not his underlying ability to discriminate *S*1 stimuli from *S*2 stimuli, the properties of the ROC curve as a whole should be

[4] These category names are more intuitive when thinking of *S*1 and *S*2 as "signal absent" and "signal present." Then a hit is a successful detection of the signal, a miss is a failure to detect the signal, a correct rejection is an accurate assessment that no signal was presented, and a false alarm is a detection of a signal where none existed.

[5] Since hit rate and miss rate sum to 1, miss rate does not provide any extra information beyond that provided by hit rate and can be ignored; similarly for false alarm rate and correct rejection rate.

[6] Note that the example ROC curve in Fig. 3.1b is depicted as having been constructed from confidence data (Fig. 3.1a), rather than from direct experimental manipulations on the observer's criterion for responding "*S*2". See the section titled *Constructing pseudo type 1 ROC curves from type 2 data* below.

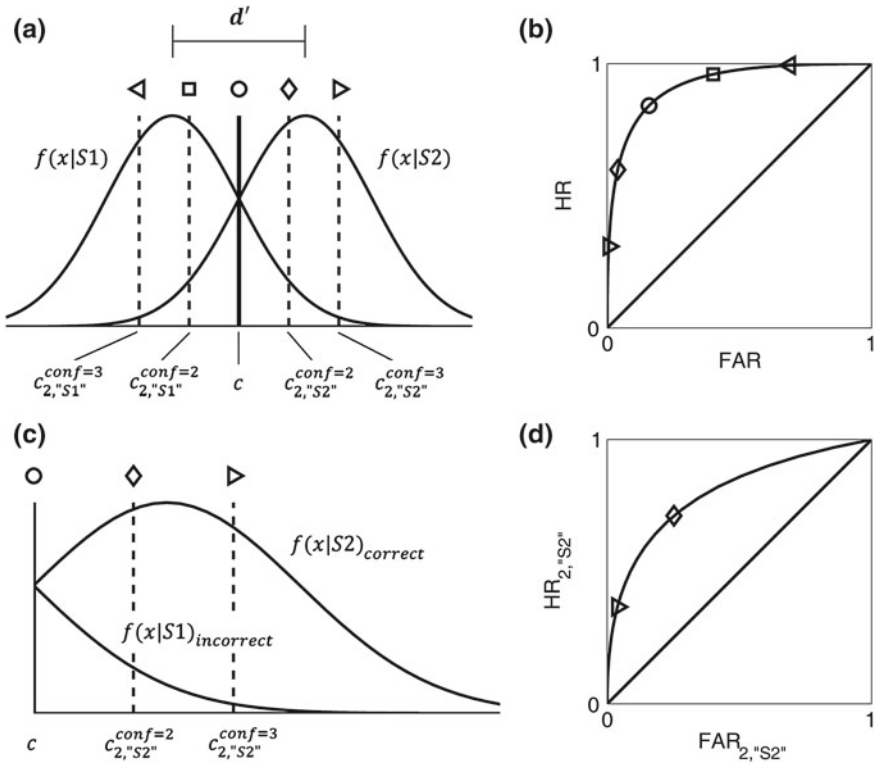


**Fig. 3.1** Signal detection theory models of type 1 and type 2 ROC curves. **a** *Type 1 SDT model*. On each trial, a stimulus generates an internal response $x$ within an observer, who must use $x$ to decide whether the stimulus was $S1$ or $S2$. For each stimulus type, $x$ is drawn from a normal distribution. The distance between these distributions is $d'$, which measures the observer's ability to discriminate $S1$ from $S2$. The stimulus is classified as "$S2$" if $x$ exceeds a decision criterion $c$, and "$S1$" otherwise. In this example, the observer also rates decision confidence on a scale of 1–3 by comparing $x$ to the additional response-specific type 2 criteria (*dashed vertical lines*). **b** *Type 1 ROC curve*. $d'$ and $c$ determine false alarm rate (*FAR*) and hit rate (*HR*). By holding $d'$ constant and changing $c$, a characteristic set of (FAR, HR) points—the ROC curve—can be generated. In this example, shapes on the ROC curve mark the (FAR, HR) generated when using the corresponding criterion in panel **a** to classify the stimulus. (Note that, because this type 1 ROC curve is generated in part by the type 2 criteria in panel 1**a**, it is actually a pseudo type 1 ROC curve, as discussed later in this paper.) **c** *Type 2 task for "S2" responses*. Consider only the trials where the observer classifies the stimulus as "$S2$," i.e. only the portion of the graph in panel a exceeding $c$. Then the $S2$ stimulus distribution corresponds to correct trials, and the $S1$ distribution to incorrect trials. The placement of the type 2 criteria determines the probability of high confidence for correct and incorrect trials—type 2 HR and type 2 FAR. $d'$ and $c$ jointly determine to what extent correct and incorrect trials for each response type are distinguishable. **d** *Type 2 ROC curve for "S2" responses*. The distributions in panel c can be used to derive type 2 FAR and HR for "$S2$" responses. By holding $d'$ and $c$ constant and changing $c_{2,"S2"}$, a set of type 2 (FAR, HR) points for "$S2$" responses—a response-specific type 2 ROC curve—can be generated. In this example, shapes on the ROC curve mark the ($FAR_{2,"S2"}$, $HR_{2,"S2"}$) generated when using the corresponding criterion in panel **c** to rate confidence

informative regarding the observer's *sensitivity* in discriminating *S*1 from *S*2, independent of the observer's overall *response bias* for producing "*S*2" responses. The observer's sensitivity thus determines the set of possible (FAR, HR) pairs the observer can produce (i.e. the ROC curve), whereas the observer's response bias determines which amongst those possible pairs is actually exhibited, depending on whether the observer is conservative or liberal in responding "*S*2." Higher sensitivity is associated with greater area underneath the ROC curve, whereas more conservative response bias is associated with (FAR, HR) points falling more towards the lower-left portion of the ROC curve.

Measures of task performance have implied ROC curves [12, 19]. An implied ROC curve for a given measure of performance is a set of (FAR, HR) pairs that yield the same value for the measure. Thus, to the extent that empirical ROC curves dissociate sensitivity from bias, they provide an empirical target for theoretical measures of performance to emulate. If a proposed measure of sensitivity does not have implied ROC curves that match the properties of empirical ROC curves, then this measure cannot be said to provide a bias-free measure of sensitivity.

A core empirical strength of SDT ([10, 12]; Fig. 3.1a) is that it provides a simple computational model that provides close fits to empirical ROC curves [10, 20]. According to SDT, the observer performs the task of discriminating *S*1 from *S*2 by evaluating internal responses along a decision axis. Every time an *S*1 stimulus is shown, it produces in the mind of the observer an internal response drawn from a Gaussian probability density function. *S*2 stimulus presentations also generate such normally distributed internal responses. For the sake of simplicity, in the following we will assume that the probability density functions for *S*1 and *S*2 have an equal standard deviation $\sigma$.

The observer is able to discriminate *S*1 from *S*2 just to the extent that the internal responses produced by these stimuli are distinguishable, such that better sensitivity for discriminating *S*1 from *S*2 is associated with larger separation between the *S*1 and *S*2 internal response distributions. The SDT measure of sensitivity, $d'$, is thus the distance between the means of the *S*1 and *S*2 distributions, measured in units of their common standard deviation:

$$d' = \frac{\mu_{S2} - \mu_{S1}}{\sigma}$$

By convention, the internal response where the *S*1 and *S*2 distributions intersect is defined to have the value of zero, so that $\mu_{S2} = \sigma\, d'/2$ and $\mu_{S1} = -\sigma\, d'/2$. For simplicity, and without loss of generality, we can set $\sigma = 1$.

In order to classify an internal response $x$ on a given trial as originating from an *S*1 or *S*2 stimulus, the observer compares the internal response to a *decision criterion*, $c$, and only produces "*S*2" classifications for internal responses that surpass the criterion.

$$\text{response} = \begin{cases} \text{“}S1\text{”}, & x \leq c \\ \text{“}S2\text{”}, & x > c \end{cases}$$

Since hit rate is the probability of responding "*S*2" when an *S*2 stimulus is shown, it can be calculated on the SDT model as the area underneath the portion of the *S*2 probability density function that exceeds $c$. Since the cumulative distribution function for the normal distribution with mean $\mu$ and standard deviation $\sigma$ evaluated at $x$ is

$$\Phi(x, \mu, \sigma) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

then hit rate can be derived from the parameters of the SDT model as

$$\text{HR} = 1 - \Phi(c, \mu_{S2}) = 1 - \Phi\left(c, \frac{d'}{2}\right)$$

And similarly,

$$\text{FAR} = 1 - \Phi(c, \mu_{S1}) = 1 - \Phi\left(c, -\frac{d'}{2}\right)$$

where omitting the $\sigma$ parameter in $\phi$ is understood to be equivalent to setting $\sigma = 1$.

By systematically altering the value of $c$ while holding $d'$ constant, a set of (FAR, HR) pairs ranging between (0, 0) and (1, 1) can be generated, tracing out the shape of the ROC curve (Fig. 3.1b). The family of ROC curves predicted by SDT matches well with empirical ROC curves across a range of experimental tasks and conditions [10, 20].

The parameters of the SDT model can be recovered from a given (FAR, HR) pair as

$$d' = z(\text{HR}) - z(\text{FAR})$$
$$c = -0.5 \times [z(\text{HR}) + z(\text{FAR})]$$

where $z$ is the inverse of the normal cumulative distribution function. Thus, SDT analysis allows us to separately characterize an observer's sensitivity ($d'$) and response bias ($c$) on the basis of a single (FAR, HR) pair, obviating the need to collect an entire empirical ROC curve in order to separately characterize sensitivity and bias—provided that the assumptions of the SDT model hold.

### *3.2.2 Type 2 SDT*

Suppose we extend the empirical task described above, such that after classifying the stimulus as *S*1 or *S*2, the observer must provide a confidence rating that characterizes the likelihood of the stimulus classification being correct. This confidence rating task can be viewed as a secondary discrimination task. Just as the observer first had to discriminate whether the stimulus was *S*1 or *S*2 by means of

providing a stimulus classification response, the observer now must discriminate whether that stimulus classification response itself was correct or incorrect by means of providing a confidence rating.[7] Following convention, we will refer to the task of classifying the stimulus as the "type 1" task, and the task of classifying the accuracy of the stimulus classification as the "type 2" task [2, 9].

#### 3.2.2.1 Type 2 Hit Rates and False Alarm Rates

A similar set of principles for the analysis of the type 1 task may be applied to the type 2 task. Consider the simple case where the observer rates confidence as either "high" or "low." We can then distinguish 4 possible outcomes in the type 2 task: high confidence correct trials, low confidence correct trials, low confidence incorrect trials, and high confidence incorrect trials. By direct analogy with the type 1 analysis, we may refer to these outcomes as type 2 hits, type 2 misses, type 2 correct rejections, and type 2 false alarms, respectively (Table 3.2).[8]

Type 2 hit rate and type 2 false alarm rate summarize an observer's type 2 performance and may be calculated as

$$\text{type 2 HR} = \text{HR}_2 = p(\text{high conf} \mid \text{stim} = \text{resp}) = \frac{n(\text{high conf correct})}{n(\text{correct})}$$

$$\text{type 2 FAR} = \text{FAR}_2 = p(\text{high conf} \mid \text{stim} \neq \text{resp}) = \frac{n(\text{high conf incorrect})}{n(\text{incorrect})}$$

Since the binary classification task we have been discussing has two kinds of correct trials (hits and correct rejections) and two kinds of incorrect trials (misses and false alarms), the classification of type 2 performance can be further subdivided into a *response-specific* analysis, where we consider type 2 performance only for trials where the type 1 stimulus classification response was "*S*1" or "*S*2" (Table 3.3).[9]

---

[7] In principle, since the observer should always choose the stimulus classification response that is deemed most likely to be correct, then in a two-alternative task he should always judge that the chosen response is more likely to be correct than it is to be incorrect. Intuitively, then, the type 2 decision actually consists in deciding whether the type 1 response is *likely* to be correct or not, where the standard for what level of confidence merits being labeled as "likely to be correct" is determined by a subjective criterion than can be either conservative or liberal. Nonetheless, viewing the type 2 task as a discrimination between correct and incorrect stimulus classifications facilitates comparison with the type 1 task.

[8] The analogy is more intuitive when thinking of *S*1 as "signal absent" and *S*2 as "signal present". Then the type 2 analogue of "signal absent" is an incorrect stimulus classification, whereas the analogue of "signal present" is a correct stimulus classification. The type 2 task can then be thought of as involving the detection of this type 2 "signal."

[9] It is also possible to conduct a stimulus-specific analysis and construct stimulus-specific type 2 ROC curves. For *S*1 stimuli, this would consist in a plot of p(high conf|correct rejection) vs p(high conf|false alarm). Likewise for *S*2 stimuli—p(high conf|hit) vs p(high conf|miss). However, as will be made clear later in the text, the present approach to analyzing type 2 ROC

**Table 3.2** Possible outcomes for the type 2 task

| Accuracy | Confidence | |
|---|---|---|
| | Low | High |
| Incorrect | Type 2 correct rejection | Type 2 false alarm |
| Correct | Type 2 miss | Type 2 hit |

**Table 3.3** Possible outcomes for the type 2 task, contingent on type 1 response (i.e., response-specific type 2 outcomes)

| Response | | | Confidence | |
|---|---|---|---|---|
| | | | Low | High |
| "*S*1" | Accuracy | Incorrect (Type 1 miss) | $CR_{2,\text{"S1"}}$ | $FA_{2,\text{"S1"}}$ |
| | | Correct (Type 1 correct rejection) | $Miss_{2,\text{"S1"}}$ | $Hit_{2,\text{"S1"}}$ |
| "*S*2" | Accuracy | Incorrect (Type 1 false alarm) | $CR_{2,\text{"S2"}}$ | $FA_{2,\text{"S2"}}$ |
| | | Correct (Type 1 hit) | $Miss_{2,\text{"S2"}}$ | $Hit_{2,\text{"S2"}}$ |

Thus, when considering type 2 performance only for "*S*1" responses,

$$\text{HR}_{2,\text{"S1"}} = p(\text{high conf} \mid \text{stim} = S1, \text{resp} = \text{"S1"}) = \frac{n(\text{high conf correct rejection})}{n(\text{correct rejection})}$$

$$\text{FAR}_{2,\text{"S1"}} = p(\text{high conf} \mid \text{stim} = S2, \text{resp} = \text{"S1"}) = \frac{n(\text{high conf miss})}{n(\text{miss})}$$

where the subscript "*S*1" indicates that these are type 2 data for type 1 "*S*1" responses.

Similarly for "*S*2" responses,

(Footnote 9 continued)
curves in terms of the type 1 SDT model requires each type 2 (FAR, HR) pair to be generated by the application of a type 2 criterion to two overlapping distributions. For stimulus-specific type 2 data, the corresponding type 1 model consists of only one stimulus distribution, with separate type 2 criteria for "*S*1" and "*S*2" responses generating the type 2 FAR and type 2 HR. (e.g. for the *S*2 stimulus, a type 2 criterion for "*S*1" responses rates confidence for type 1 misses, and a separate type 2 criterion for "*S*2" responses rates confidence for type 1 hits.) Thus there is no analogue of meta-$d'$ for stimulus-specific type 2 data, since $d'$ is only defined with respect to the relationship between two stimulus distributions, whereas stimulus-specific analysis is restricted to only one stimulus distribution. It is possible that an analysis of stimulus-specific type 2 ROC curves could be conducted by positing how the type 2 criteria on either side of the type 1 criterion are coordinated, or similarly by supposing that the observer rates confidence according to an overall type 2 decision variable. For more elaboration, see the section below titled "Comparison of the current approach to that of [9]."

$$\mathrm{HR}_{2,\text{“}S2\text{”}} = p(\text{high conf} \mid \text{stim} = S2,\ \text{resp} = \text{“}S2\text{”}) = \frac{n(\text{high conf hit})}{n(\text{hit})}$$

$$\mathrm{FAR}_{2,\text{“}S2\text{”}} = p(\text{high conf} \mid \text{stim} = S1,\ \text{resp} = \text{“}S2\text{”}) = \frac{n(\text{high conf false alarm})}{n(\text{false alarm})}$$

From the above definitions, it follows that overall type 2 FAR and HR are weighted averages of the response-specific type 2 FARs and HRs, where the weights are determined by the proportion of correct and incorrect trials originating from each response type:

$$\begin{aligned}\mathrm{HR}_2 &= \frac{n(\text{high conf correct})}{n(\text{correct})} = \frac{n(\text{high conf hit}) + n(\text{high conf CR})}{n(\text{hit}) + n(\text{CR})} \\ &= \frac{n(\text{hit}) \times \mathrm{HR}_{2,\text{“}S2\text{”}} + n(\text{CR}) \times \mathrm{HR}_{2,\text{“}S1\text{”}}}{n(\text{hit}) + n(\text{CR})} \\ &= p(\text{hit} \mid \text{correct}) \times \mathrm{HR}_{2,\text{“}S2\text{”}} + [1 - p(\text{hit} \mid \text{correct})] \times \mathrm{HR}_{2,\text{“}S1\text{”}}\end{aligned}$$

And similarly,

$$\mathrm{FAR}_2 = p(\text{FA} \mid \text{incorrect}) \times \mathrm{FAR}_{2,\text{“}S2\text{”}} + [1 - p(\text{FA} \mid \text{incorrect})] \times \mathrm{FAR}_{2,\text{“}S1\text{”}}$$

Confidence rating data may be richer than a mere binary classification. In the general case, the observer may rate confidence on either a discrete or continuous scale ranging from 1 to $H$. In this case, we can arbitrarily select a value $h$, $1 < h \leq H$, such that all confidence ratings greater than or equal to $h$ are classified as “high confidence” and all others, “low confidence.” We can denote this choice of imposing a binary classification upon the confidence data by writing e.g. $H_2^{\text{conf}=h}$, where the superscript conf $= h$ indicates that this type 2 hit rate was calculated using a classification scheme where $h$ was the smallest confidence rating considered to be “high.” Thus, for instance,

$$\mathrm{HR}_{2,\text{“}S2\text{”}}^{\text{conf}=h} = p(\text{high conf} \mid \text{stim} = S2,\ \text{resp} = \text{“}S2\text{”}) = p(\text{conf} \geq h \mid \text{hit})$$

Each choice of $h$ generates a type 2 (FAR, HR) pair, and so calculating these for multiple values of $h$ allows for the construction of a type 2 ROC curve with multiple points. When using a discrete confidence rating scale ranging from 1 to $H$, there are $H - 1$ ways of selecting $h$, allowing for the construction of a type 2 ROC curve with $H - 1$ points.

#### 3.2.2.2 Adding Response-Specific Type 2 Criteria to the Type 1 SDT Model to Capture Type 2 Data

As with the type 1 task, type 2 ROC curves allow us to separately assess an observer’s sensitivity (how well confidence ratings discriminate correct from incorrect trials) and response bias (the overall propensity for reporting high

confidence) in the type 2 task. However, fitting a computational model to type 2 ROC curves is somewhat more complicated than in the type 1 case. It is not appropriate to assume that correct and incorrect trials are associated with normal probability density functions in a direct analogy to the *S*1 and *S*2 distributions of type 1 SDT. The reason for this is that specifying the parameters of the type 1 SDT model—$d'$ and $c$—places strong constraints on the probability density functions for correct and incorrect trials, and these derived distributions are not normally distributed [9]. In addition to this theoretical consideration, it has also been empirically demonstrated that conducting a type 2 SDT analysis that assumes normal distributions for correct and incorrect trials does not give a good fit to data [6].

Thus, the structure of the SDT model for type 2 performance must take into account the structure of the SDT model for type 1 performance. Galvin et al. [9] presented an approach for the SDT analysis of type 2 data based on analytically deriving formulae for the type 2 probability density functions under a suitable transformation of the type 1 decision axis. Here we present a simpler alternative approach on the basis of which response-specific type 2 ROC curves can be derived directly from the type 1 model.

In order for the type 1 SDT model to characterize type 2 data, we first need an added mechanism whereby confidence ratings can be generated. This can be accomplished by supposing that the observer simply uses additional decision criteria, analogous to the type 1 criterion $c$, to generate a confidence rating on the basis of the internal response $x$ on a given trial. In the simplest case, the observer makes a binary confidence rating—high or low—and thus needs to use two additional decision criteria to rate confidence for each kind of type 1 response. Call these response-specific type 2 criteria $c_{2,\text{“}S1\text{”}}$ and $c_{2,\text{“}S2\text{”}}$, where $c_{2,\text{“}S1\text{”}} < c$ and $c_{2,\text{“}S2\text{”}} > c$. Intuitively, confidence increases as the internal response $x$ becomes more distant from $c$, i.e. as the internal response becomes more likely to have been generated by one of the two stimulus distributions.[10] More formally,

$$\text{confidence}_{\text{resp}=\text{“}S1\text{”}} = \begin{cases} \text{low}, & x \geq c_{2,\text{“}S1\text{”}} \\ \text{high}, & x < c_{2,\text{“}S1\text{”}} \end{cases}$$

$$\text{confidence}_{\text{resp}=\text{“}S2\text{”}} = \begin{cases} \text{low}, & x \leq c_{2,\text{“}S2\text{”}} \\ \text{high}, & x > c_{2,\text{“}S2\text{”}} \end{cases}$$

In the more general case of a discrete confidence scale ranging from 1 to $H$, then $H - 1$ type 2 criteria are required to rate confidence for each response type. (See e.g. Fig. 3.1a, where two type 2 criteria on left/right of the type 1 criterion allow for confidence for “*S*1”/“*S*2” responses to be rated on a scale of 1–3.) We may define

[10] See “Comparison of the current approach to that of Galvin et al. [9]” and footnote 12 for a more detailed consideration of the type 2 decision axis.

$$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = \left(c_{2,\text{“}S1\text{”}}^{\text{conf}=2}, c_{2,\text{“}S1\text{”}}^{\text{conf}=3}, \ldots, c_{2,\text{“}S1\text{”}}^{\text{conf}=H}\right)$$
$$\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = \left(c_{2,\text{“}S2\text{”}}^{\text{conf}=2}, c_{2,\text{“}S2\text{”}}^{\text{conf}=3}, \ldots, c_{2,\text{“}S2\text{”}}^{\text{conf}=H}\right)$$

where e.g. $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}}$ is a tuple containing the $H-1$ type 2 criteria for “$S1$” responses. Each $c_{2,\text{“}S1\text{”}}^{\text{conf}=y}$ denotes the type 2 criterion such that internal responses more extreme (i.e. more distant from the type 1 criterion) than $c_{2,\text{“}S1\text{”}}^{\text{conf}=y}$ are associated with confidence ratings of at least $y$. More specifically,

$$\text{confidence}_{\text{resp}=\text{“}S1\text{”}} = \begin{cases} 1, & x \geq c_{2,\text{“}S1\text{”}}^{\text{conf}=2} \\ y, & c_{2,\text{“}S1\text{”}}^{\text{conf}=y+1} \leq x < c_{2,\text{“}S1\text{”}}^{\text{conf}=y}, \quad 1 < y < H \\ H, & x < c_{2,\text{“}S1\text{”}}^{\text{conf}=H} \end{cases}$$
$$\text{confidence}_{\text{resp}=\text{“}S2\text{”}} = \begin{cases} 1, & x \leq c_{2,\text{“}S2\text{”}}^{\text{conf}=2} \\ y, & c_{2,\text{“}S2\text{”}}^{\text{conf}=y} < x \leq c_{2,\text{“}S2\text{”}}^{\text{conf}=y+1}, \quad 1 < y < H \\ H, & x > c_{2,\text{“}S2\text{”}}^{\text{conf}=H} \end{cases}$$

The type 1 and type 2 decision criteria must have a certain ordering in order for the SDT model to be meaningful. Response-specific type 2 criteria corresponding to higher confidence ratings must be more distant from $c$ than type 2 criteria corresponding to lower confidence ratings. Additionally, $c$ must be larger than all type 2 criteria for “$S1$” responses but smaller than all type 2 criteria for “$S2$” responses. For convenience, we may define

$$\underline{\boldsymbol{c}}_{\text{ascending}} = \left(c_{2,\text{“}S1\text{”}}^{\text{conf}=H}, c_{2,\text{“}S1\text{”}}^{\text{conf}=H-1}, \ldots, c_{2,\text{“}S1\text{”}}^{\text{conf}=1}, c, c_{2,\text{“}S2\text{”}}^{\text{conf}=1}, c_{2,\text{“}S2\text{”}}^{\text{conf}=2}, \ldots, c_{2,\text{“}S2\text{”}}^{\text{conf}=H}\right)$$

The ordering of decision criteria in $\underline{\boldsymbol{c}}_{\text{ascending}}$ from first to last is the same as the ordering of the criteria from left to right when displayed on an SDT graph (e.g. Fig. 3.1a). These decision criteria are properly ordered only if each element of $\underline{\boldsymbol{c}}_{\text{ascending}}$ is at least as large as the previous element, i.e. only if the Boolean function $\gamma\left(\underline{\boldsymbol{c}}_{\text{ascending}}\right)$ defined below is true:

$$\gamma\left(\underline{\boldsymbol{c}}_{\text{ascending}}\right) = \bigcap_{i=1}^{2H-2} \underline{\boldsymbol{c}}_{\text{ascending}}(i+1) \geq \underline{\boldsymbol{c}}_{\text{ascending}}(i)$$

It will be necessary to use this function later on when discussing how to fit SDT models to type 2 data.

#### 3.2.2.3 Calculating Response-Specific Type 2 (FAR, HR) from the Type 1 SDT Model with Response-Specific Type 2 Criteria

Now let us consider how to calculate response-specific type 2 HR and type 2 FAR from the type 1 SDT model. Recall that

$$\mathrm{HR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h} = p(\mathrm{conf} \geq h \mid \mathrm{stim} = S2,\ \mathrm{resp} = \text{“}S2\text{”}) = \frac{p(\mathrm{conf} \geq h,\ \mathrm{hit})}{p(\mathrm{hit})}$$

As discussed above, $p(\mathrm{hit})$, the hit rate, is the probability that an $S2$ stimulus generates an internal response that exceeds the type 1 criterion $c$. Similarly, $p(\mathrm{conf} \geq h, \mathrm{hit})$, the probability of a hit endorsed with high confidence, is just the probability that an $S2$ stimulus generates an internal response that exceeds the high-confidence type 2 criterion for "$S2$" responses, $c_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h}$. Thus, we can straightforwardly characterize the probabilities in the numerator and denominator of $\mathrm{HR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h}$ in terms of the type 1 SDT parameters, as follows:

$$\mathrm{HR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h} = \frac{p(\mathrm{conf} \geq h,\ \mathrm{hit})}{p(\mathrm{hit})} = \frac{1 - \Phi\left(c_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h}, \frac{d'}{2}\right)}{1 - \Phi\left(c, \frac{d'}{2}\right)}$$

By similar reasoning,

$$\mathrm{FAR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h} = \frac{1 - \Phi\left(c_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h}, -\frac{d'}{2}\right)}{1 - \Phi\left(c, -\frac{d'}{2}\right)}$$

And likewise for "$S1$" responses,

$$\mathrm{HR}_{2,\text{“}S1\text{”}}^{\mathrm{conf}=h} = \frac{\Phi\left(c_{2,\text{“}S1\text{”}}^{\mathrm{conf}=h}, -\frac{d'}{2}\right)}{\Phi\left(c, -\frac{d'}{2}\right)}$$

$$\mathrm{FAR}_{2,\text{“}S1\text{”}}^{\mathrm{conf}=h} = \frac{\Phi\left(c_{2,\text{“}S1\text{”}}^{\mathrm{conf}=h}, \frac{d'}{2}\right)}{\Phi\left(c, \frac{d'}{2}\right)}$$

Figure 3.1c illustrates how type 2 (FAR, HR) arise from type 1 $d'$ and $c$ along with a type 2 criterion. For instance, suppose $h = 3$. Then the type 2 hit rate for "$S2$" responses, $\mathrm{HR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=3}$, is the probability of a high confidence hit (the area in the $S2$ distribution beyond $c_{2,\text{“}S2\text{”}}^{\mathrm{conf}=3}$) divided by the probability of a hit (the area in the $S2$ distribution beyond $c$).

By systematically altering the value of the type 2 criteria while holding $d'$ and $c$ constant, a set of ($\mathrm{FAR}_2$, $\mathrm{HR}_2$) pairs ranging between (0, 0) and (1, 1) can be generated, tracing out a curvilinear prediction for the shape of the type 2 ROC curve (Fig. 3.1d). Thus, according to this SDT account, specifying type 1

sensitivity ($d'$) and response bias ($c$) is already sufficient to determine response-specific type 2 sensitivity (i.e. the family of response-specific type 2 ROC curves).

### 3.2.3 Comparison of the Current Approach to that of Galvin et al. [9]

Before continuing with our treatment of SDT analysis of type 2 data, we will make some comparisons between this approach and the one described in Galvin et al. [9].

#### 3.2.3.1 SDT Approaches to Type 2 Performance

Galvin et al. were concerned with characterizing the *overall* type 2 ROC curve, rather than response-specific type 2 ROC curves. On their modeling approach, an ($FAR_2$, $HR_2$) pair can be generated by setting a single type 2 criterion on a type 2 decision axis. All internal responses that exceed this type 2 criterion are labeled "high confidence," and all others "low confidence." By systematically changing the location of this type 2 criterion on the decision axis, the entire overall type 2 ROC curve can be traced out.

However, if the internal response $x$ is used to make the binary confidence decision in this way, the ensuing type 2 ROC curve behaves oddly, typically containing regions where it extends below the line of chance performance [9]. This suboptimal behavior is not surprising, in that comparing the raw value of $x$ to a single criterion value essentially recapitulates the decision rule used in the type 1 task and does not take into account the relationship between $x$ and the observer's type 1 criterion, which is crucial for evaluating type 1 performance. The solution is that some *transformation* of $x$ must be used as the type 2 decision variable, ideally one that depends upon both $x$ and $c$.

For instance, consider the transformation $t(x) = |x - c|$. This converts the initial raw value of the internal response, $x$, into the distance of $x$ from the type 1 criterion. This transformed value can then plausibly be compared to a single type 2 criterion to rate confidence, e.g. an observer might rate confidence as high whenever $t(x) > 1$. Other transformations for the type 2 decision variable are possible, and the choice is not arbitrary, since different choices for type 2 decision variables can lead to different predictions for the type 2 ROC curve [9]. The optimal type 2 ROC curve (i.e. the one that maximizes area under the curve) is derived by using the likelihood ratio of the type 2 probability density functions as the type 2 decision variable [9, 10].

We have adopted a different approach thus far. Rather than characterizing an overall ($FAR_2$, $HR_2$) pair as arising from the comparison of a single type 2 decision variable to a single type 2 criterion, we have focused on response-specific ($FAR_2$, $HR_2$) data arising from comparisons of the type 1 internal response $x$ to

separate type 2 decision criteria for "*S*1" and "*S*2" responses (e.g. Fig. 3.1a). Thus, our approach would characterize the overall ($FAR_2$, $HR_2$) as arising from a pair of response-specific type 2 criteria set on either side of the type 1 criterion on the type 1 decision axis, rather than from a single type 2 criterion set on a type 2 decision axis. We have posited no constraints on the setting of these type 2 criteria other than that they stand in appropriate ordinal relationships to eachother. For the sake of brevity in comparing these two approaches, in the following we will refer to Galvin et al.'s approach as G and the current approach as C.

#### 3.2.3.2 Type 2 Decision Rules and Response-Specific Type 2 Criterion Setting

Notice that choosing a reasonable type 2 decision variable for G is equivalent to setting constraints on the relationship between type 2 criteria for "*S*1" and "*S*2" responses on C. For instance, on *G* suppose that the type 2 decision variable is defined as $t(x) = |x - c|$ and confidence is high if $t(x) > 1$. On C, this is equivalent to setting response-specific type 2 criteria symmetrically about the type 1 criterion, i.e. $t(c_{2,\text{“}S1\text{”}}) = t(c_{2,\text{“}S2\text{”}}) = |c_{2,\text{“}S1\text{”}} - c| = |c_{2,\text{“}S2\text{”}} - c| = 1$. In other words, assuming (on G) the general rule that confidence is high whenever the distance between $x$ and $c$ exceeds 1 requires (on C) that the type 2 criteria for each response type both satisfy this property of being 1 unit away from $c$. Any other way of setting the type 2 criteria for C would yield outcomes inconsistent with the decision rule posited by G. Similarly, if the type 2 decision rule is that confidence is high when type 2 likelihood ratio $LR_2(x) > c_{LR2}$, this same rule on C would require $LR_2(c_{2,\text{“}S1\text{”}}) = LR_2(c_{2,\text{“}S2\text{”}}) = c_{LR2}$, i.e. that type 2 criteria for both response types be set at the locations of $x$ on either side of $c$ corresponding to a type 2 likelihood ratio of $c_{LR2}$.

On G, choosing a suboptimal type 2 decision variable can lead to decreased area under the overall type 2 ROC curve. This can be understood on C as being related to the influence of response-specific type 2 criterion placement on the response-specific type 2 (FAR, HR) points, which in turn affect the overall type 2 (FAR, HR) points. As shown above, overall type 2 FAR and HR are weighted averages of the corresponding response-specific type 2 FARs and HRs. But computing a weighted average for two (FAR, HR) pairs on a concave down ROC curve will yield a new (FAR, HR) pair that lies below the original ROC curve. As a consequence, more exaggerated differences in the response-specific type 2 FAR and HR due to more exaggerated difference in response-specific type 2 criterion placement will tend to drive down the area below the overall type 2 ROC curve. Thus, the overall type 2 ROC curve may decrease even while the response-specific curves stay constant, depending on how criterion setting for each response type is coordinated. This reduced area under the overall type 2 ROC curve on C due to response-specific type 2 criterion placement is closely related to reduced area under the overall type 2 ROC curve on G due to choosing a suboptimal type 2 decision variable.

For example, consider the SDT model where $d' = 2$, $c = 0$, $c_{2,\text{“}S1\text{”}} = -1$, and $c_{2,\text{“}S2\text{”}} = 1$. This model yields $\text{FAR}_{2,\text{“}S1\text{”}} = \text{FAR}_{2,\text{“}S2\text{”}} = \text{FAR}_2 = 0.14$ and $\text{HR}_{2,\text{“}S1\text{”}} = \text{HR}_{2,\text{“}S2\text{”}} = \text{HR}_2 = 0.59$. The type 1 criterion is optimally placed and the type 2 criteria are symmetrically placed around it. This arrangement of criteria on C turns out to be equivalent to using the type 2 likelihood ratio on G, and thus yields an optimal type 2 performance. Now consider the SDT model where $d' = 2$, $c = 0$, $c_{2,\text{“}S1\text{”}} = -1.5$, and $c_{2,\text{“}S2\text{”}} = 0.76$. This model yields $\text{FAR}_{2,\text{“}S1\text{”}} = 0.04$, $\text{HR}_{2,\text{“}S1\text{”}} = 0.37$, $\text{FAR}_{2,\text{“}S2\text{”}} = 0.25$, $\text{HR}_{2,\text{“}S2\text{”}} = 0.71$, and overall $\text{FAR}_2 = 0.14$, $\text{HR}_2 = 0.54$. Although $d'$ and $c$ are the same as in the previous example, now the type 2 criteria are set asymmetrically about $c$, yielding different outcomes for the type 2 FAR and HR for “*S*1” and “*S*2” responses. This has the effect of yielding a lower overall $\text{HR}_2$ (0.54 vs. 0.59) in spite of happening to yield the same $\text{FAR}_2$ (0.14). Thus, this asymmetric arrangement of response-specific type 2 criteria yields worse performance on the overall type 2 ROC curve than the symmetric case for the same values of $d'$ and $c$. On G, this can be understood as being the result of choosing a suboptimal type 2 decision variable in the second example (i.e. a decision variable that is consistent with the way the response-specific type 2 criteria have been defined on C). In this case, the asymmetric placement of the response-specific type 2 criteria is inconsistent with a type 2 decision variable based on the type 2 likelihood ratio.

#### 3.2.3.3 A Method for Assessing Overall Type 2 Sensitivity Based on the Approach of Galvin et al.

In the upcoming section, we will discuss our methodology for quantifying type 2 sensitivity with meta-$d'$. Meta-$d'$ essentially provides a single measure that jointly characterizes the areas under the response-specific type 2 ROC curves for both “*S*1” and “*S*2” responses, and in this way provides a measure of overall type 2 sensitivity. However, in doing so, it treats the relationships of type 2 criteria across response types as purely a matter of criterion setting. However, as we have discussed, coordination of type 2 criterion setting could also be seen as arising from the construction of a type 2 decision variable, where the choice of decision variable influences area under the overall type 2 ROC curve. We take it to be a substantive conceptual, and perhaps empirical, question as to whether it is preferable to characterize these effects as a matter of criterion setting (coordinating response-specific type 2 criteria) or sensitivity (constructing a type 2 decision variable). However, if one were to decide that for some purpose it were better to view this as a sensitivity effect, then the characterization of type 2 performance provided by Galvin et al. may be preferable to that of the current approach.

In the interest of recognizing this, we provide free Matlab code available online (see note at the end of the manuscript) that implements one way of using Galvin et al.'s approach to evaluate an observer's overall type 2 performance. Given the

parameters of an SDT model, this code outputs the theoretically optimal[11] overall type 2 ROC curve—i.e. the overall type 2 ROC curve based on type 2 likelihood ratio, which has the maximum possible area under the curve. Maniscalco and Lau [13], building on the suggestions of Galvin et al. [9], proposed that one way of evaluating an observer's type 2 performance is to compare her empirical type 2 ROC curve with the theoretical type 2 ROC curve, given her type 1 performance. By comparing an observer's empirical overall type 2 ROC curve with the theoretically optimal overall type 2 ROC curve based on type 2 likelihood ratios, the observer's overall type 2 sensitivity can be assessed with respect to the SDT-optimal level. This approach will capture potential variation in area under the overall type 2 ROC curve that is ignored (treated as a response-specific criterion effect) by the meta-$d'$ approach.

#### 3.2.3.4 Advantages of the Current Approach

Our SDT treatment of type 2 performance has certain advantages over that of Galvin et al. One advantage is that it does not require making an explicit assumption regarding what overall type 2 decision variable an observer uses, or even that the observer constructs such an overall type 2 decision variable to begin with.[12] This is because our approach allows the type 2 criteria for each response to vary independently, rather than positing a fixed relationship between their locations. Thus, if an observer does construct an overall type 2 decision variable, our treatment will capture this implicitly by means of the relationship between the response-specific type 2 criteria; and if an observer does not use an overall type 2 decision variable to begin with, our treatment can accommodate this behavior. The question of what overall type 2 decision variables, if any, observers tend to use is a substantive empirical question, and so it is preferable to avoid making assumptions on this matter if possible.

A second, related advantage is that our approach is potentially more flexible than Galvin et al.'s in capturing the behavior of response-specific type 2 ROC curves, without loss of flexibility in capturing the overall type 2 ROC curve. (Since overall type 2 ROC curves depend on the response-specific curves, as shown above, our focus on characterizing the response-specific curves does not entail a deficit in capturing the overall curve.) A third advantage is that our approach provides a simple way to derive response-specific type 2 ROC curves from the

[11] Provided the assumptions of the SDT model are correct.

[12] Of course, our approach must at least implicitly assume a type 2 decision variable *within* each response type. In our treatment, the implicit type 2 decision variable for each response type is just the distance of $x$ from $c$. However, for the analysis of response-specific type 2 performance for the equal variance SDT model, distance from criterion and type 2 likelihood ratio are equivalent decision variables. This is because they vary monotonically with each other [9], and so produce the same type 2 ROC curve [5, 21].

type 1 SDT model, whereas deriving the overall type 2 ROC curve is more complex under Galvin et al.'s approach and depends upon the type 2 decision variable being assumed.

## 3.3 Characterizing Type 2 Sensitivity in Terms of Type 1 SDT: Meta-$d'$

Since response-specific type 2 ROC curves can be derived directly from $d'$ and $c$ on the SDT model, this entails a tight theoretical relationship between type 1 and type 2 performance. One practical consequence is that type 2 sensitivity—the empirical type 2 ROC curves—can be quantified in terms of the type 1 SDT parameters $d'$ and $c$ [13]. However, it is necessary to explicitly differentiate instances when $d'$ is meant to characterize type 1 performance from those instances when $d'$ (along with $c$) is meant to characterize type 2 performance. Here we adopt the convention of using the variable names meta-$d'$ and meta-$c$ to refer to type 1 SDT parameters when used to characterize type 2 performance. We will refer to the type 1 SDT model as a whole, when used to characterize type 2 performance, as the meta-SDT model. Essentially, $d'$ and $c$ describe the type 1 SDT model fit to the type 1 ROC curve,[13] whereas meta-$d'$ and meta-$c$—the meta-SDT model—quantify the type 1 SDT model when used exclusively to fit type 2 ROC curves.

How do we go about using the type 1 SDT model to quantify type 2 performance? There are several choices to make before a concrete method can be proposed. In the course of discussing these issues, we will put forth the methodological approach originally proposed by Maniscalco and Lau [13].

### *3.3.1 Which Type 2 ROC Curves?*

As discussed in the preceding section "Comparison of the current approach to that of Galvin et al. [9]," we find the meta-SDT fit that provides the best simultaneous fit to the response-specific type 2 ROC curves for "$S1$" and "$S2$" responses, rather than finding a model that directly fits the overall type 2 ROC curve. As explained in more detail in that prior discussion, we make this selection primarily because (1) it allows more flexibility and accuracy in fitting the overall data set, and (2) it does not require making an explicit assumption regarding what type 2 decision variable the observer might use for confidence rating.

[13] When the multiple points on the type 1 ROC curve are obtained using confidence rating data, it is arguably preferable to calculate $d'$ and $c$ only from the (FAR, HR) pair generated purely by the observer's type 1 response. The remaining type 1 ROC points incorporate confidence rating data and depend on type 2 sensitivity, and so estimating $d'$ on the basis of these ROC points may confound type 1 and type 2 sensitivity. See the section below titled "Response-specific meta-$d'$ and the unequal variance SDT model".

### 3.3.2 Which Way of Using Meta-d′ and Meta-c to Derive Response-Specific Type 2 ROC Curves?

A second consideration is how to characterize the response-specific type 2 ROC curves using meta-$d'$ and meta-$c$. For the sake of simplifying the analysis, and for the sake of facilitating comparison between $d'$ and meta-$d'$, an appealing option is to *a priori* fix the value of meta-$c$ so as to be similar to the empirically observed type 1 response bias $c$, thus effectively allowing meta-$d'$ to be the sole free parameter that characterizes type 2 sensitivity. However, since there are multiple ways of measuring type 1 response bias [12], there are also multiple ways of fixing the value of meta-$c$ on the basis of $c$. In addition to the already-introduced $c$, type 1 response bias can be measured with the relative criterion, $c'$:

$$c' = c/d'$$

This measure takes into account how extreme the criterion is, *relative to* the stimulus distributions.

Bias can also be measured as $\beta$, the ratio of the probability density function for $S2$ stimuli to that of $S1$ stimuli at the location of the decision criterion:

$$\beta = e^{cd'}$$

Figure 3.2 shows an example of how $c$, $c'$, and $\beta$ relate to the stimulus distributions when bias is fixed and $d'$ varies. Panel a shows an SDT diagram for $d' = 3$ and $c = 1$. In panel b, $d' = 1$ and the three decision criteria are generated by setting $c$, $c'$, and $\beta$ to the equivalent values of those exhibited by these measures in panel a. Arguably, $c'$ performs best in terms of achieving a similar "cut" between the stimulus distributions in panels a and b. This is an intuitive result given that $c'$ essentially adjusts the location of $c$ according to $d'$. Thus, holding $c'$ constant ensures that, as $d'$ changes, the location of the decision criterion remains in a similar location with respect to the means of the two stimulus distributions.

By choosing $c'$ as the measure of response bias that will be held constant in the estimation of meta-$d'$, we can say that when the SDT and meta-SDT models are fit to the same data set, they will have similar type 1 response bias, in the sense that they have the same $c'$ value. This in turn allows us to interpret a subject's meta-$d'$ in the following way: "Suppose there is an ideal subject whose behavior is perfectly described by SDT, and who performs this task with a similar level of response bias (i.e. same $c'$) as the actual subject. Then in order for our ideal subject to produce the actual subject's response-specific type 2 ROC curves, she would need her $d'$ to be equal to meta-$d'$."

Thus, meta-$d'$ can be found by fitting the type 1 SDT model to response-specific type 2 ROC curves, with the constraint that meta-$c' = c'$. (Note that in the below we list meta-$c$, rather than meta-$c'$, as a parameter of the meta-SDT model. The constraint meta-$c' = c'$ can thus be satisfied by ensuring meta-$c$ = meta-$d' \times c'$.)

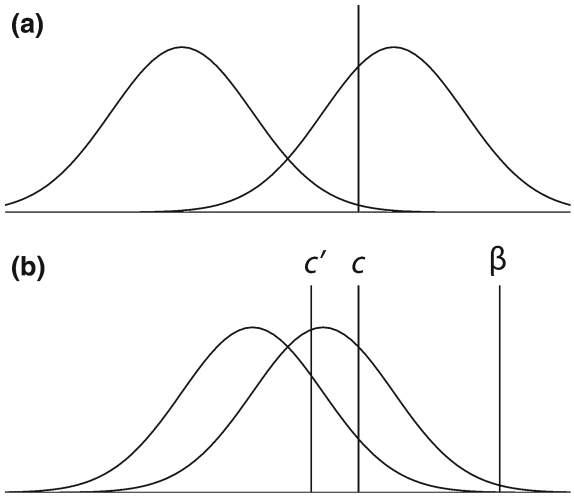


**Fig. 3.2** Example behavior of holding response bias constant as $d'$ changes for $c$, $c'$, and $\beta$. **a** An SDT graph where $d' = 3$ and $c = 1$. The criterion location can also be quantified as $c' = c/d' = 1/3$ and $\log \beta = c \times d' = 3$. **b** An SDT graph where $d' = 1$. The three decision criteria plotted here represent the locations of the criteria that preserve the value of the corresponding response bias exhibited in panel a. So e.g. the criterion marked $c'$ in panel b has the same value of $c'$ as the criterion in panel a (=1/3), and likewise for $c$ (constant value of 1) and $\beta$ (constant value of 3)

### *3.3.3 What Computational Method of Fitting?*

If the response-specific type 2 ROC curves contain more than one empirical ($FAR_2$, $HR_2$) pair, then in general an exact fit of the model to the data is not possible. In this case, fitting the model to the data requires minimizing some loss function, or maximizing some metric of goodness of fit.

Here we consider the procedure for finding the parameters of the type 1 SDT model that maximize the likelihood of the response-specific type 2 data. Maximum likelihood approaches for fitting SDT models to type 1 ROC curves with multiple data points have been established [4, 16]. Here we adapt these existing type 1 approaches to the type 2 case. The likelihood of the type 2 data can be characterized using the multinomial model as

$$L_{\text{type 2}}(\theta \mid \text{data}) \propto \prod_{y,s,r} \text{Prob}_\theta(\text{conf} = y \mid \text{stim} = s,\ \text{resp} = r)^{n_{\text{data}}(\text{conf}=y \mid \text{stim}=s,\ \text{resp}=r)}$$

Maximizing likelihood is equivalent to maximizing log-likelihood, and in practice it is typically more convenient to work with log-likelihoods. The log-likelihood for type 2 data is given by

$$\log\ L_{\text{type 2}}(\theta \mid \text{data}) \propto \sum_{y,s,r} n_{\text{data}} \log\ \text{Prob}_\theta$$

$\theta$ is the set of parameters for the meta-SDT model:

$$\theta = \left(\text{meta-}d', \text{meta-}c, \text{meta-}\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}}, \text{meta-}\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}}\right)$$

$n_{\text{data}}(\text{conf} = y \mid \text{stim} = s,\ \text{resp} = r)$ is a count of the number of times in the data a confidence rating of $y$ was provided when the stimulus was $s$ and response was $r$. $y$, $s$, and $r$ are indices ranging over all possible confidence ratings, stimulus classes, and stimulus classification responses, respectively.
$\text{prob}_\theta(\text{conf} = y \mid \text{stim} = s,\ \text{resp} = r)$ is the model-predicted probability of generating confidence rating $y$ for trials where the stimulus and response were $s$ and $r$, given the parameter values specified in $\theta$.

Calculation of these type 2 probabilities from the type 1 SDT model is similar to the procedure used to calculate the response-specific type 2 FAR and HR. For notational convenience, below we express these probabilities in terms of the standard SDT model parameters, omitting the “meta” prefix.

For convenience, define

$$\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S1\text{”}} = \left(c, c_{2,\text{“}S1\text{”}}^{\text{conf}=2}, c_{2,\text{“}S1\text{”}}^{\text{conf}=3}, \ldots, c_{2,\text{“}S1\text{”}}^{\text{conf}=H}, -\infty\right)$$
$$\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S2\text{”}} = \left(c, c_{2,\text{“}S2\text{”}}^{\text{conf}=2}, c_{2,\text{“}S2\text{”}}^{\text{conf}=3}, \ldots, c_{2,\text{“}S2\text{”}}^{\text{conf}=H}, \infty\right)$$

Then

$$\text{Prob}(\text{conf} = y \mid \text{stim} = S1,\ \text{resp} = \text{“}S1\text{”}) = \frac{\Phi\left(\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S1\text{”}}(y), -\frac{d'}{2}\right) - \Phi\left(\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S1\text{”}}(y+1), -\frac{d'}{2}\right)}{\Phi\left(c, -\frac{d'}{2}\right)}$$

$$\text{Prob}(\text{conf} = y \mid \text{stim} = S1,\ \text{resp} = \text{“}S2\text{”}) = \frac{\Phi\left(\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S2\text{”}}(y+1), -\frac{d'}{2}\right) - \Phi\left(\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S2\text{”}}(y), -\frac{d'}{2}\right)}{1 - \Phi\left(c, -\frac{d'}{2}\right)}$$

$$\text{Prob}(\text{conf} = y \mid \text{stim} = S2,\ \text{resp} = \text{“}S2\text{”}) = \frac{\Phi\left(\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S2\text{”}}(y+1), \frac{d'}{2}\right) - \Phi\left(\dot{\underline{\boldsymbol{c}}}_{2,\text{“}S2\text{”}}(y), \frac{d'}{2}\right)}{1 - \Phi\left(c, \frac{d'}{2}\right)}$$

An illustration of how these type 2 probabilities are derived from the type 1 SDT model is provided in Fig. 3.3.

The multinomial model used as the basis for calculating likelihood treats each discrete type 2 outcome ($\text{conf} = y \mid \text{stim} = s, \text{resp} = r$) as an event with a fixed probability that occurred a certain number of times in the data set, where outcomes across trials are assumed to be statistically independent. The probability of the entire set of type 2 outcomes across all trials is then proportional to the product of the probability of each individual type 2 outcome, just as e.g. the probability of

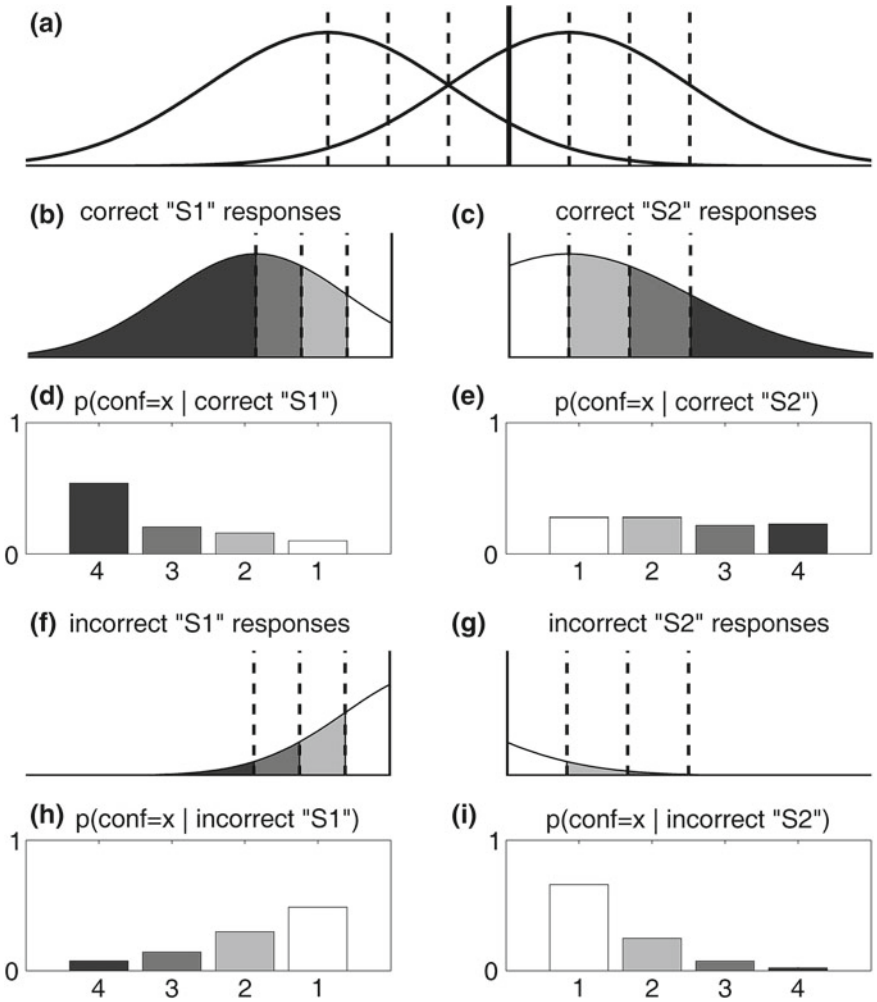


**Fig. 3.3** Type 2 response probabilities from the SDT model. **a** An SDT graph with $d' = 2$ and decision criteria $c = 0.5$, $\boldsymbol{c}_{2,\text{“}S1\text{”}} = (0, -0.5, -1)$, and $\boldsymbol{c}_{2,\text{“}S2\text{”}} = (1, 1.5, 2)$. The type 1 criterion (solid vertical line) is set to the value of 0.5, corresponding to a conservative bias for providing “*S*2” responses, in order to create an asymmetry between “*S*1” and “*S*2” responses for the sake of illustration. Seven decision criteria are used in all, segmenting the decision axis into 8 regions. Each region corresponds to one of the possible permutations of type 1 and type 2 responses, as there are two possible stimulus classifications and four possible confidence ratings. **b–i** Deriving probability of confidence rating contingent on type 1 response and accuracy. How would the SDT model depicted in panel (**a**) predict the probability of each confidence rating for correct “*S*1” responses? Since we wish to characterize “*S*1” responses, we need consider only the portion of the SDT graph falling to the left of the type 1 criterion. Since “*S*1” responses are only correct when the *S*1 stimulus was actually presented, we can further limit our consideration to internal responses generated by *S*1 stimuli. This is depicted in panel (**b**). This distribution is further subdivided into 4 levels of confidence by the 3 type 2 criteria (*dashed vertical lines*), where darker regions correspond to higher confidence. The area under the *S*1 curve in each of these

◀ regions, divided by the total area under the *S*1 curve that falls below the type 1 criterion, yields the probability of reporting each confidence level, given that the observer provided a correct "*S*1" response. Panel (**d**) shows these probabilities as derived from areas under the curve in panel (**b**). The remaining panels display the analogous logic for deriving confidence probabilities for incorrect "*S*1" responses (**f**, **h**), correct "*S*2" responses (**c**, **e**), and incorrect "*S*2" responses (**g**, **i**)

throwing 4 heads and 6 tails for a fair coin is proportional to $0.5^4 \times 0.5^6$. (Calculation of the exact probability depends on a combinatorial term which is invariant with respect to $\theta$ and can therefore be ignored for the purposes of maximum likelihood fitting.)

Likelihood, $L(\theta)$, can be thought of as measuring how probable the empirical data is, according to the model parameterized with $\theta$. A very low $L(\theta)$ indicates that the model with $\theta$ would be very unlikely to generate a pattern like that observed in the data. A higher $L(\theta)$ indicates that the data are more in line with the typical behavior of data produced by the model with $\theta$. Mathematical optimization techniques can be used to find the values of $\theta$ that maximize the likelihood, i.e. that create maximal concordance between the empirical distribution of outcomes and the model-expected distribution of outcomes.

The preceding approach for quantifying type 2 sensitivity with the type 1 SDT model—i.e. for fitting the meta-SDT model—can be summarized as a mathematical optimization problem:

$$\theta^* = \arg\max_{\theta} L_{\text{type 2}}(\theta \mid \text{data}), \quad \text{subject to:} \quad \text{meta-}c' = c',\ \gamma\left(\text{meta-}\underline{\boldsymbol{c}}_{\text{ascending}}\right)$$

where type 2 sensitivity is quantified by meta-$d' \in \theta^*$.

$\gamma\left(\text{meta-}\underline{\boldsymbol{c}}_{\text{ascending}}\right)$ is the Boolean function described previously, which returns a value of "true" only if the type 1 and type 2 criteria stand in appropriate ordinal relationships.

We provide free Matlab code, available online, for implementing this maximum likelihood procedure for fitting the meta-SDT model to a data set (see note at the end of the manuscript).

### *3.3.4 Toy Example of Meta-*$\mathbf{d}'$ *Fitting*

An illustration of the meta-$d'$ fitting procedure is demonstrated in Fig. 3.4 using simulated data. In this simulation, we make the usual SDT assumption that on each trial, presentation of stimulus S generates an internal response $x$ that is drawn from the probability density function of S, and that a type 1 response is made by comparing $x$ to the decision criterion $c$. However, we now add an extra mechanism to the model to allow for the possibility of added noise in the type 2 task. Let us call the internal response used to rate confidence $x_2$. The type 1 SDT model we

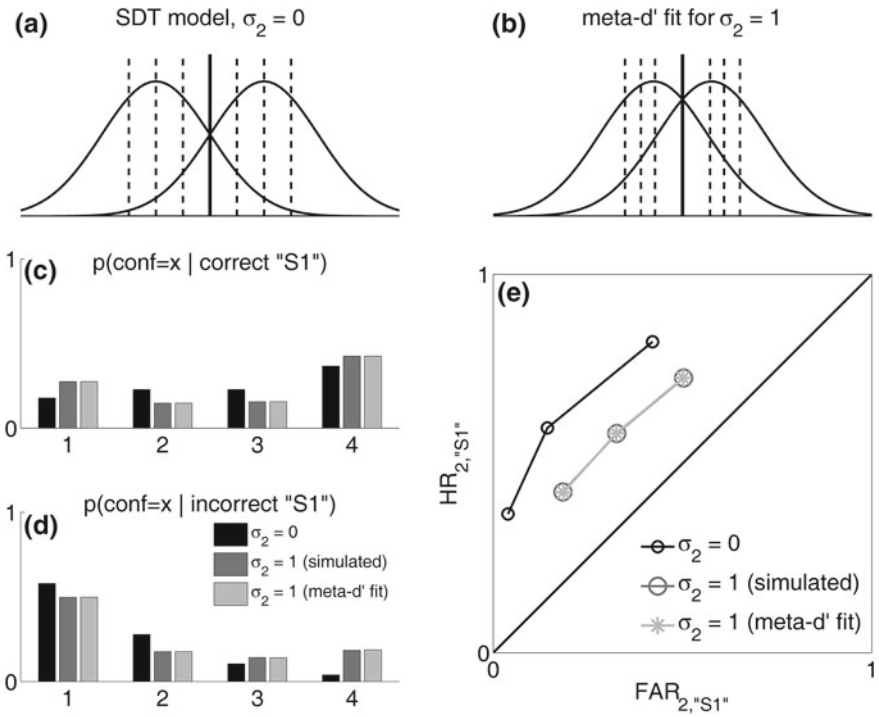


**Fig. 3.4** Fitting meta-$d'$ to response-specific type 2 data. **a** Graph for the SDT model where $d' = 2$ and $\sigma_2 = 0$ (see text for details). **b** A model identical to that in panel a, with the exception that $\sigma_2 = 1$, was used to create simulated data. This panel displays the SDT graph of the parameters for the meta-$d'$ fit to the $\sigma_2 = 1$ data. **c**, **d** Response-specific type 2 probabilities. The maximum likelihood method of fitting meta-$d'$ to type 2 data uses response-specific type 2 probabilities as the fundamental unit of analysis. The type 1 SDT parameters that maximize the likelihood of the type 2 data yield distributions of response-specific type 2 probabilities closely approximating the empirical (here, simulated) distributions. Here we only show the probabilities for "*S*1" responses; because of the symmetry of the generating model, "*S*2" responses follow identical distributions. **e** Response-specific type 2 ROC curves. ROC curves provide a more informative visualization of the type 2 data than the raw probabilities. Here it is evident that there is considerably less area under the type 2 ROC curve for the $\sigma_2 = 1$ simulation than is predicted by the $\sigma_2 = 0$ model. The meta-$d'$ fit provides a close match to the simulated data

have thus far considered assumes $x_2 = x$. In this example, we suppose that $x_2$ is a noisier facsimile of $x$. Formally,

$$x_2 = x + \xi, \quad \xi \sim N(0, \sigma_2)$$

where $N(0, \sigma_2)$ is the normal distribution with mean 0 and standard deviation $\sigma_2$. The parameter $\sigma_2$ thus determines how much noisier $x_2$ is than $x$. For $\sigma_2 = 0$ we expect meta-$d' = d'$, and for $\sigma_2 > 0$ we expect meta-$d' < d'$.

The simulated observer rates confidence on a 4-point scale by comparing $x_2$ to response-specific type 2 criteria, using the previously defined decision rules for confidence in the type 1 SDT model.[14]

We first considered the SDT model with $d' = 2$, $c = 0$, $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.5, -1, -1.5)$, $\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.5, 1, 1.5)$ and $\sigma_2 = 0$. Because $\sigma_2 = 0$, this is equivalent to the standard type 1 SDT model. The SDT graph for these parameter values is plotted in Fig. 3.4a. Using these parameter settings, we computed the theoretical probability of each confidence rating for each permutation of stimulus and response. These probabilities for “*S*1” responses are shown in panels c and d, and the corresponding type 2 ROC curve is shown in panel e. (Because the type 1 criterion $c$ is unbiased and the type 2 criteria are set symmetrically about $c$, confidence data for “*S*2” responses follow an identical distribution to that of “*S*1” responses and are not shown.)

Next we simulated 10,000,000 trials using the same parameter values as the previously considered model, with the exception that $\sigma_2 = 1$. With this additional noise in the type 2 task, type 2 sensitivity should decrease. This decrease in type 2 sensitivity can be seen in the type 2 ROC curve in panel e. There is more area underneath the type 2 ROC curve when $\sigma_2 = 0$ than when $\sigma_2 = 1$.

We performed a maximum likelihood fit of meta-$d'$ to the simulated type 2 data using the fmincon function in the optimization toolbox for Matlab (MathWorks, Natick, MA), yielding a fit with parameter values meta-$d' = 1.07$, meta-$c = 0$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.51, -0.77, -1.06)$, and meta-$\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.51, 0.77, 1.06)$. The SDT graph for these parameter values is plotted in Fig. 3.4b.

Panels c and d demonstrate the component type 2 probabilities used for computing the type 2 likelihood. The response-specific type 2 probabilities for $\sigma_2 = 0$ are not distributed the same way as those for $\sigma_2 = 1$, reflecting the influence of adding noise to the internal response for the type 2 task. Computing meta-$d'$ for the $\sigma_2 = 1$ data consists in finding the parameter values of the ordinary type 1 SDT model that maximize the likelihood of the $\sigma_2 = 1$ response-specific type 2 data. This results in a type 1 SDT model whose theoretical type 2 probabilities closely

[14] Note that for this model, it is possible for $x$ and $x_2$ to be on opposite sides of the type 1 decision criterion $c$ (see, e.g. Fig. 3.5a, b). This is not problematic, since only $x$ is used to provide the type 1 stimulus classification. It is also possible for $x_2$ to surpass some of the type 2 criteria on the opposite side of $c$. For instance, suppose that $x = -0.5$, $x_2 = +0.6$, $c = 0$, and $c_{2,\text{“}S2\text{”}}^{\text{conf}=h} = +0.5$. Then $x$ is classified as an *S*1 stimulus, and yet $x_2$ surpasses the criterion for rating “*S*2” responses with a confidence of $h$. Thus, there is potential for the paradoxical result whereby the type 1 response is “*S*1” and yet the type 2 confidence rating is rated highly due to the relatively strong “*S*2”-ness of $x_2$. In this example, the paradox is resolved by the definition of the type 2 decision rules stated above, which stipulate that internal responses are only evaluated with respect to the response-specific type 2 criteria that are congruent with the type 1 response. Thus, in this case, the decision rule would not compare $x_2$ with the type 2 criteria for “*S*2” responses to begin with. Instead, it would find that $x_2$ does not surpass the minimal confidence criterion for “*S*1” responses (i.e., $x_2 > c > c_{2,\text{“}S1\text{”}}^{\text{conf}=2}$) and would therefore assign $x_2$ a confidence of 1. Thus, in this case, the paradoxical outcome is averted. But such potentially paradoxical results need to be taken into account for any SDT model that posits a potential dissociation between $x$ and $x_2$.

match the empirical type 2 probabilities for the simulated $\sigma_2 = 1$ data (Fig. 3.4c, d). Because type 2 ROC curves are closely related to these type 2 probabilities, the meta-$d'$ fit also produces a type 2 ROC curve closely resembling the simulated curve, as shown in panel e.

### 3.3.5 *Interpretation of Meta-*$\mathbf{d'}$

Notice that because meta-$d'$ characterizes type 2 sensitivity purely in terms of the type 1 SDT model, it does not explicitly posit any mechanisms by means of which type 2 sensitivity varies. Although the meta-$d'$ fitting procedure gave a good fit to data simulated by the toy $\sigma_2$ model discussed above, it could also produce similarly good fits to data generated by different models that posit completely different mechanisms for variation in type 2 performance. In this sense, meta-$d'$ is descriptive but not explanatory. It describes how an ideal SDT observer with similar type 1 response bias as the actual subject would have achieved the observed type 2 performance, rather than explain how the actual subject achieved their type 2 performance.

The primary virtue of using meta-$d'$ is that it allows us to quantify type 2 sensitivity in a principled SDT framework, and compare this against SDT expectations of what type 2 performance *should have been*, given performance on the type 1 task, all while remaining agnostic about the underlying processes. For instance, if we find that a subject has $d' = 2$ and meta-$d' = 1$, then (1) we have taken appropriate SDT-inspired measures to factor out the influence of response bias in our measure of type 2 sensitivity; (2) we have discovered a violation of the SDT expectation that meta-$d' = d' = 2$, giving us a point of reference in interpreting the subject's metacognitive performance in relation to their primary task performance and suggesting that the subject's metacognition is suboptimal (provided the assumptions of the SDT model hold); and (3) we have done so while making minimal assumptions and commitments regarding the underlying processes.

Another important point for interpretation concerns the raw meta-$d'$ value, as opposed to its value in relation to $d'$. Suppose observers A and B both have meta-$d' = 1$, but $d'_{\mathrm{A}} = 1$ and $d'_{\mathrm{B}} = 2$. Then there is a sense in which they have equivalent metacognition, as their confidence ratings are equally sensitive in discerning correct from incorrect trials. But there is also a sense in which A has superior metacognition, since A was able to achieve the same level of meta-$d'$ as B in spite of a lower $d'$. In a sense, A is more metacognitively ideal, according to SDT. We can refer to the first kind of metacognition, which depends only on meta-$d'$, as "absolute type 2 sensitivity," and the second kind, which depends on the relationship between meta-$d'$ and $d'$, as "relative type 2 sensitivity." Absolute and relative type 2 sensitivity are distinct constructs that inform us about distinct aspects of metacognitive performance.

For more on the interpretation of meta-$d'$, see Maniscalco and Lau [13].

## 3.4 Response-Specific Meta-$d'$

Thus far we have considered how to characterize an observer's overall type 2 sensitivity using meta-$d'$, expounding upon the method originally introduced in Maniscalco and Lau [13]. Here we show how to extend this analysis and characterize response-specific type 2 sensitivity in terms of the type 1 SDT model.

In the below we focus on "*S*1" responses, but similar considerations apply for "*S*2" responses.

We wish to find the type 1 SDT parameters $\theta$ that provide the best fit to the type 2 ROC curve for "*S*1" responses, i.e. the set of empirical $\left(\text{FAR}_{2,\text{“}S1\text{”}}^{\text{conf}=h}, \text{HR}_{2,\text{“}S1\text{”}}^{\text{conf}=h}\right)$ for all $h$ satisfying $2 \leq h \leq \text{H}$. Thus, we wish to find the $\theta$ that maximizes the likelihood of the type 2 probabilities for "*S*1" responses, using the usual meta-$d'$ fitting approach. This is essentially equivalent to applying the original meta-$d'$ procedure described above to the subset of the model and data pertaining to "*S*1" responses.

Thus, we wish to solve the optimization problem

$$\theta_{\text{“}S1\text{”}}^{*} = \underset{\theta_{\text{“}S1\text{”}}}{\arg\max}\, L_{2,\text{“}S1\text{”}}(\theta_{\text{“}S1\text{”}} \mid \text{data}),$$

$$\text{subject to:} \quad \text{meta-}c'_{\text{“}S1\text{”}} = c', \quad \gamma\left(\text{meta-}\underline{\boldsymbol{c}}_{\text{ascending}}\right)$$

where

$$\theta_{\text{“}S1\text{”}} = \left(\text{meta-}d'_{\text{“}S1\text{”}}, \text{meta-}c_{\text{“}S1\text{”}}, \text{meta-}\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}}\right)$$

$$L_{2,\text{“}S1\text{”}}(\theta_{\text{“}S1\text{”}} \mid \text{data}) \propto \prod_{y,s} \text{Prob}_{\theta}(\text{conf} = y \mid \text{stim} = s,\ \text{resp} = \text{“}S1\text{”})^{n_{\text{data}}(\text{conf}=y \mid \text{stim}=s,\ \text{resp}=\text{“}S1\text{”})}$$

meta-$d'_{\text{“}S1\text{”}} \in \theta_{\text{“}S1\text{”}}^{*}$ measures type 2 sensitivity for "*S*1" responses.

The differences between this approach and the "overall" meta-$d'$ fit are straightforward. The same likelihood function is used, but with the index $r$ fixed to the value "*S*1". $\theta_{\text{“}S1\text{”}}$ is equivalent to $\theta$ except for its omission of meta$\underline{c}_{2,\text{“}S2\text{”}}$, since type 2 criteria for "*S*2" responses are irrelevant for fitting "*S*1" type 2 ROC curves. The type 1 criterion meta-$c_{\text{“}S1\text{”}}$ is listed with a "*S*1" subscript to distinguish it from meta-$c_{\text{“}S2\text{”}}$, the type 1 criterion value from the maximum likelihood fit to "*S*2" type 2 data. Since the maximum likelihood fitting procedure enforces the constraint meta-$c'_{\text{“}S1\text{”}} = c'$, it follows that meta-$c_{\text{“}S1\text{”}} = \text{meta-}d'_{\text{“}S1\text{”}} \times c'$. Thus, in the general case where meta-$d'_{\text{“}S1\text{”}} \neq \text{meta-}d'_{\text{“}S2\text{”}}$ and $c' \neq 0$, it is also true that meta-$c_{\text{“}S1\text{”}} \neq \text{meta-}c_{\text{“}S2\text{”}}$.

We provide free Matlab code, available online, for implementing this maximum likelihood procedure for fitting the response-specific meta-SDT model to a data set (see note at the end of the manuscript).

### 3.4.1 Toy Example of Response-Specific Meta-d′ Fitting

An illustration of the response-specific meta-$d'$ fitting procedure is demonstrated in Fig. 3.5 using simulated data. We use a similar model as that used in the previous toy example of meta-$d'$ fitting. That is, we use the usual type 1 SDT model, except we suppose that the internal response used to produce the type 2 judgment, $x_2$, may be a noisier version of its type 1 counterpart, $x$. This time, we additionally allow the degree of added noisiness in $x_2$ to differ for "$S1$" and "$S2$" responses. Formally,

$$x_2 = \begin{cases} x + \xi_{\text{“}S1\text{”}}, & \xi_{\text{“}S1\text{”}} \sim N(0, \sigma_{2\,\text{“}S1\text{”}}) \quad \text{if } x \leq c \\ x + \xi_{\text{“}S2\text{”}}, & \xi_{\text{“}S2\text{”}} \sim N\left(0, \sigma_{2,\text{“}S2\text{”}}\right) \quad \text{if } x > c \end{cases}$$

Different levels of type 2 noisiness for each response type allows for the possibility that response-specific type 2 sensitivity can differ for "$S1$" and "$S2$" responses.

We first considered the SDT model with $d' = 2, c = 0, \underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.5, -1, -1.5)$, $\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.5, 1, 1.5)$ and $\sigma_{2,\text{“}S1\text{”}} = \sigma_{2,\text{“}S2\text{”}} = 0$. Because $\sigma_{2,\text{“}S1\text{”}} = \sigma_{2,\text{“}S2\text{”}} = 0$, this is equivalent to the standard type 1 SDT model. The SDT graph for these parameter values were used in the previous example, as illustrated in Fig. 3.4a. Using these parameter settings, we constructed theoretical response-specific type 2 ROC curves, as shown in Fig. 3.5c, d.

Next we simulated 10,000,000 trials using the same parameter values as the previously considered model, with the exception that $\sigma_{2,\text{“}S1\text{”}} = 2/3$ and $\sigma_{2,\text{“}S2\text{”}} = 1/3$. Since $\sigma_{2,\text{“}S2\text{”}} < \sigma_{2,\text{“}S1\text{”}}$, for these simulated data there is more area underneath the type 2 ROC curve for "$S2$" than for "$S1$" responses (Fig. 3.5c, d). The simulated distributions of $x_2$ values for correct and incorrect "$S1$" and "$S2$" responses is shown in the top halves of Fig. 3.5a, b. Note that this model generates some $x_2$ values that lie on the opposite side of the type 1 criterion as the corresponding $x$ value (which determines the type 1 response). For all such trials, the type 1 response was determined only by $x$ and confidence was set to 1. See footnote 14 above for more details.

We first performed a maximum likelihood fit of overall meta-$d'$ to the simulated data, yielding a fit with parameter values meta-$d' = 1.17$, meta-$c = 0$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.59, -0.79, -1.01)$, and meta-$\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.43, 0.80, 1.2)$.The theoretical type 2 ROC curves predicted by the SDT model with these parameter values is displayed in Fig. 3.5c, d alongside the simulated data. Inspection of these graphs suggests that the meta-$d'$ fit was able to account for differences in overall levels of confidence for "$S1$" and "$S2$" responses, as reflected by the fact that the response-specific curves are scaled in such a way as to mirror the scaling of the empirical type 2 ROC curves. However, the meta-$d'$ fit cannot account for the difference in type 2 sensitivity for "$S1$" and "$S2$" responses. Instead, the fit produces overlapping type 2 ROC curves located midway between the empirical "$S1$" and "$S2$" curves, as if capturing something analogous to the average type 2 sensitivity for each response type. (See the inset of Fig. 3.5d for a plot of the meta-$d'$ type 2 ROC curves for both response types.)

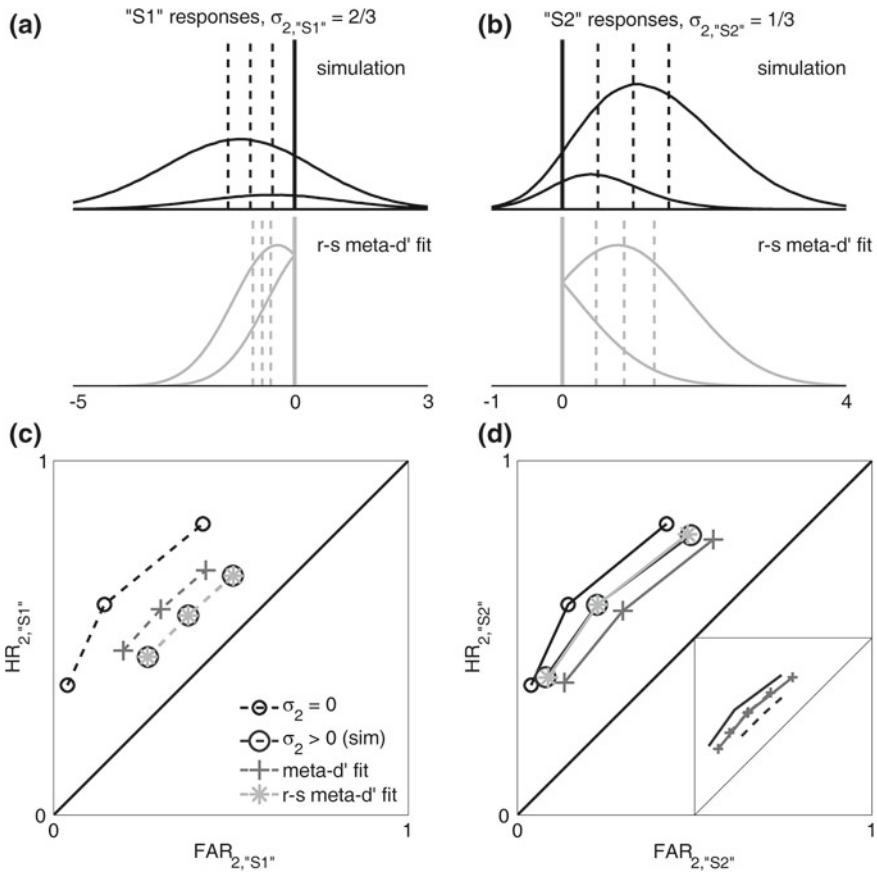


**Fig. 3.5** Response-specific meta-$d'$ fitting. **a** Simulated data and meta-$d'$ fit for "S1" responses. *Top* Simulated distribution of $x_2$ values for correct and incorrect "*S*1" responses for simulated data with $\sigma_{2,\text{“}S1\text{”}} = 2/3$. (See main text for details.) Note that many $x_2$ values initially labeled "*S*1" cross over to the other side of the type 1 criterion after having type 2 noise added. These are considered to be "*S*1" responses with confidence =1. See footnote 14 in main text for further discussion. *Bottom* SDT parameters of meta-$d'_{\text{“}S1\text{”}}$ fit. **b** Same as A, but for "*S*2" responses. **c** Type 2 ROC curves for "S1" responses. Setting $\sigma_{2,\text{“}S1\text{”}} = 2/3$ substantially reduces type 2 sensitivity, as revealed by the comparison of area under the ROC curves for $\sigma_{2,\text{“}S1\text{”}} = 2/3$ and $\sigma_{2,\text{“}S1\text{”}} = 0$. Response-specific meta-$d'$ fits the data well, but meta-$d'$ provides an overestimate. **d** Type 2 ROC curves for "S2" responses. Response-specific meta-$d'$ fits the "*S*2" data well, but meta-$d'$ provides an underestimate. *Inset* Type 2 ROC curves for both "*S*1" and "*S*2" responses, shown for the simulated data (*black*) and the meta-$d'$ fit (*gray*). The meta-$d'$ fit generates type 2 ROC curves intermediate between the empirical (simulated) "*S*1" and "*S*2" curves

Next we performed a maximum likelihood fit for response-specific meta-$d'$ to the simulated data. This yielded a fit with parameter values meta-$d'_{\text{“}S1\text{”}} = 0.77$, meta-$c_{\text{“}S1\text{”}} = 0$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.54, -0.73, -0.94)$ for “$S1$” responses, and meta-$d'_{\text{“}S2\text{”}} = 1.56$, meta-$c_{\text{“}S2\text{”}} = 0$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.48, 0.87, 1.30)$ for “$S2$” responses. The SDT graph for these parameter values is plotted in the bottom halves of Fig. 3.5a, b. The theoretical type 2 ROC curves corresponding to these fits are displayed in Fig. 3.5c, d alongside the simulated data. It is evident that the response-specific meta-$d'$ approach provides a close fit to the simulated data.

## 3.5 Response-Specific Meta-$d'$ and the Unequal Variance SDT Model

### *3.5.1 Inferring Unequal Variance from the z-ROC Curve Slope*

Thus far we have discussed SDT models assuming that the variance of the internal response distributions for $S1$ and $S2$ stimuli have equal variance. However, it is also possible to relax this assumption and allow the variances to differ. In conventional notation, we can define an additional parameter to the type 1 SDT model, $s$:

$$s = \frac{\sigma_{S1}}{\sigma_{S2}}$$

We may refer to the SDT model parameterized with $s$ as the unequal variance SDT model, or UV-SDT. We may similarly refer to the more basic SDT model we have discussed thus far as the equal variance SDT model or EV-SDT.

UV-SDT has been shown to have advantages over EV-SDT in capturing certain data sets. The primary motivation for UV-SDT arises from the analysis of type 1 z-ROC curves. Given a set of type 1 (FAR, HR) points, a z-ROC curve may be constructed by plotting z(HR) against z(FAR), where z denotes the inverse of the cumulative distribution function for the normal distribution. That is,

$$z(p) = x, \quad \text{such that } \Phi(x, 0, 1) = p$$

According to SDT, since FAR and HR are generated by the normal cumulative distribution function evaluated at some location on the decision axis $X$, it should follow that z(FAR) and z(HR) correspond to locations on $X$. More specifically, it can be shown that z(FAR) quantifies the distance between the mean of the $S1$ distribution and the criterion used to generate that FAR, as measured in units of the standard deviation of the $S1$ distribution [and similarly for z(HR)]. That is,

$$z(\mathrm{FAR}_c) = \frac{\mu_{S1} - c}{\sigma_{S1}}, \quad \mathrm{FAR}_c = 1 - \Phi(c, \mu_{S1}, \sigma_{S1})$$
$$z(\mathrm{HR}_c) = \frac{\mu_{S2} - c}{\sigma_{S2}}, \quad \mathrm{HR}_c = 1 - \Phi(c, \mu_{S2}, \sigma_{S2})$$

The slope of the z-ROC curve for a set of ($\mathrm{FAR}_c$, $\mathrm{HR}_c$) represents how changes in $z(\mathrm{HR}_c)$ relate to changes in $z(\mathrm{FAR}_c)$. According to SDT, this is equivalent to how changes in the criterion $c$, as measured in $\sigma_{S2}$ units, are related to changes in the same quantity $c$ as measured in $\sigma_{S1}$ units, since

$$z\text{-ROC slope} = \frac{\Delta z(\mathrm{HR})}{\Delta z(\mathrm{FAR})} = \frac{\Delta c/\sigma_{S2}}{\Delta c/\sigma_{S1}} = \frac{\sigma_{S1}}{\sigma_{S2}} = s$$

### 3.5.2 *Constructing Pseudo Type 1 ROC Curves from Type 2 Data*

Under EV-SDT, where $\sigma_{S1} = \sigma_{S2}$, the z-ROC curve should therefore be linear with a slope of 1, since changing $c$ by $\delta$ units of $\sigma_{S2}$ is equivalent to changing $c$ by $\delta$ units of $\sigma_{S1}$. Under UV-SDT, the z-ROC curve should be linear with a slope of $s$, since changing $c$ by $\delta$ units of $\sigma_{S1}$ is equivalent to changing $c$ by s $\times$ $\delta$ units of $\sigma_{S2}$. Thus, empirical instances of linear z-ROC curves with non-unit slope have been taken to constitute empirical support for the UV-SDT model (e.g. [20]).

Constructing empirical type 1 ROC curves requires manipulating response bias in order to collect multiple type 1 (FAR, HR) points at the same level of sensitivity. One method of accomplishing this is to place the subject in multiple experimental conditions that tend to induce different response biases, e.g. due to different base rates of stimulus presentation or payoff structures [12, 22]. However, this method is somewhat resource intensive.

A popular alternative strategy for constructing empirical type 1 ROC curves is to use the conjunction of type 1 and type 2 judgments in order to emulate distinct type 1 judgments. For instance, suppose the observer classifies a stimulus as *S*1 or *S*2 and then rates confidence as high or low. FAR and HR are determined by how often the observer responds "*S*2." But we can also imagine that, had the subject been very conservative in responding "*S*2," he might have only done so for those trials in which he endorsed the "*S*2" response with high confidence. Thus, we can compute a second (FAR, HR) pair by provisionally treating only "high confidence *S*2" trials as "*S*2" responses. Similarly, we can emulate a liberal type 1 response bias by provisionally treating anything other than a "high confidence *S*1" response as an "*S*2" response. This procedure would thus allow us to derive 3 points on the type 1 ROC curve from a single experimental session.

Following the naming convention introduced by Galvin et al. [9], we will refer to the type 1 ROC curve constructed in this way as the pseudo type 1 ROC curve,

and the extra (FAR, HR) points generated from confidence ratings as pseudo type 1 (FAR, HR). For a discrete H-point rating scale, we can derive $2H - 1$ points on the pseudo type 1 ROC curve. In addition to the usual (FAR, HR) pair as determined by the observer's stimulus classification, we can compute new pseudo (FAR, HR) pairs from "*S*1" and "*S*2" responses at each level of confidence $h > 1$, as

$$\begin{aligned}
\mathrm{HR}_{1\sim 2,\text{“}S1\text{”}}^{\mathrm{conf}=h} &= 1 - p(\mathrm{resp} = \text{“}S1\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S2) \\
\mathrm{FAR}_{1\sim 2,\text{“}S1\text{”}}^{\mathrm{conf}=h} &= 1 - p(\mathrm{resp} = \text{“}S1\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S1)
\end{aligned}$$

$$\begin{aligned}
\mathrm{HR}_{1\sim 2,\text{“}S2\text{”}}^{\mathrm{conf}=h} &= p(\mathrm{resp} = \text{“}S2\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S2) \\
\mathrm{FAR}_{1\sim 2,\text{“}S2\text{”}}^{\mathrm{conf}=h} &= p(\mathrm{resp} = \text{“}S2\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S1)
\end{aligned}$$

The subscript "$1\sim 2$" denotes that these pseudo type 1 (FAR, HR) pairs are being treated as type 1 data in spite of having been partially constructed from type 2 decisions.

The pseudo type 1 ROC curve has a straightforward interpretation on the SDT graph. Each pseudo type 1 (FAR, HR) pair can be computed from the SDT model by using the corresponding response-specific type 2 criterion in place of the type 1 criterion in the formula for FAR and HR:

$$\begin{aligned}
\mathrm{HR}_{1\sim 2,\text{“}SX\text{”}}^{\mathrm{conf}=h} &= 1 - \Phi\left(c_{2,\text{“}SX\text{”}}^{\mathrm{conf}=h}, \mu_{S2}, \sigma_{S2}\right) \\
\mathrm{FAR}_{1\sim 2,\text{“}SX\text{”}}^{\mathrm{conf}=h} &= 1 - \Phi\left(c_{2,\text{“}SX\text{”}}^{\mathrm{conf}=h}, \mu_{S1}, \sigma_{S1}\right)
\end{aligned}$$

where "SX" denotes either "*S*1" or "*S*2." Figure 3.1a, b illustrates this principle.

### 3.5.3 Dependence of Pseudo Type 1 ROC Curves on Response-Specific Type 2 ROC Curves

However, because the pseudo type 1 (FAR, HR) points depend on both type 1 and type 2 judgments, they risk confounding type 1 and type 2 sensitivity. Indeed, we will now demonstrate that pseudo type 1 (FAR, HR) points directly depend upon type 1 and type 2 ROC data. For instance, consider the pseudo type 1 (FAR, HR) for "*S*2" responses. It follows from the definition of these that

$$\begin{aligned}
\mathrm{HR}_{1\sim 2,\text{“}S2\text{”}}^{\mathrm{conf}=h} &= p(\mathrm{resp} = \text{“}S2\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S2) \\
&= p(\mathrm{conf} \geq h \mid \mathrm{resp} = \text{“}S2\text{”},\ \mathrm{stim} = S2) \times p(\mathrm{resp} = \text{“}S2\text{”} \mid \mathrm{stim} = S2) \\
&= \mathrm{HR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h} \times \mathrm{HR}_1
\end{aligned}$$

$$\begin{aligned}\mathrm{FAR}_{1\sim 2,\text{“}S2\text{”}}^{\mathrm{conf}=h} &= p(\mathrm{resp} = \text{“}S2\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S1) \\ &= p(\mathrm{conf} \geq h \mid \mathrm{resp} = \text{“}S2\text{”},\ \mathrm{stim} = S1) \times p(\mathrm{resp} = \text{“}S2\text{”} \mid \mathrm{stim} = S1) \\ &= \mathrm{FAR}_{2,\text{“}S2\text{”}}^{\mathrm{conf}=h} \times \mathrm{FAR}_1\end{aligned}$$

Similarly for “*S*1” responses,

$$\begin{aligned}\mathrm{HR}_{1\sim 2,\text{“}S1\text{”}}^{\mathrm{conf}=h} &= 1 - p(\mathrm{resp} = \text{“}S1\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S2) \\ &= 1 - [p(\mathrm{conf} \geq h \mid \mathrm{resp} = \text{“}S1\text{”},\ \mathrm{stim} = S2) \times p(\mathrm{resp} = \text{“}S1\text{”} \mid \mathrm{stim} = S2)] \\ &= 1 - \left[\mathrm{FAR}_{2,\text{“}S1\text{”}}^{\mathrm{conf}=h} \times (1 - \mathrm{HR}_1)\right]\end{aligned}$$

$$\begin{aligned}\mathrm{FAR}_{1\sim 2,\text{“}S1\text{”}}^{\mathrm{conf}=h} &= 1 - p(\mathrm{resp} = \text{“}S1\text{”},\ \mathrm{conf} \geq h \mid \mathrm{stim} = S1) \\ &= 1 - [p(\mathrm{conf} \geq h \mid \mathrm{resp} = \text{“}S1\text{”},\ \mathrm{stim} = S1) \times p(\mathrm{resp} = \text{“}S1\text{”} \mid \mathrm{stim} = S1)] \\ &= 1 - \left[\mathrm{HR}_{2,\text{“}S1\text{”}}^{\mathrm{conf}=h} \times (1 - \mathrm{FAR}_1)\right]\end{aligned}$$

Thus, if separate cognitive mechanisms govern type 1 and type 2 judgments, then it is possible that patterns in the pseudo type 1 ROC curve reflect aspects of cognitive processing pertaining to type 2, rather than type 1, judgments. One such theoretical pattern is revealed in the case of chance type 2 responding, as discussed in Clifford et al. [3]. If an observer has chance levels of type 2 sensitivity, then confidence ratings do not differentiate between correct and incorrect trials, and so $\mathrm{HR}_2 = \mathrm{FAR}_2$. The pseudo type 1 ROC points constructed from such data would consist in a linear scaling of the “true” ($\mathrm{FAR}_1$, $\mathrm{HR}_1$) pair by some constant $k = \mathrm{HR}_2 = \mathrm{FAR}_2$. Thus, the pseudo type 1 ROC curve would consist of two line segments, one connecting (0, 0) to ($\mathrm{FAR}_1$, $\mathrm{HR}_1$) (corresponding to chance type 2 performance for “*S*2” responses), the other connecting ($\mathrm{FAR}_1$, $\mathrm{HR}_1$) to (1, 1) (corresponding to chance type 2 performance for “*S*1” responses); see Clifford et al.’s Fig. 3.2c.

Here we make the observation that pseudo type 1 z-ROC curves with non-unit slope can be generated by an EV-SDT model with differences in response-specific meta-$d'$ (hereafter, RSM-SDT). By the same token, we observe that differences in the area under response-specific type 2 ROC curves can be generated purely as a consequence of the type 1 properties of the UV-SDT model. Thus, considerable caution is warranted in making inferences about the cognitive processes that underlie patterns in type 1 and type 2 ROC curves because of the possibility of confusing the effects of different variance for type 1 distributions and different suboptimalities for response-specific metacognitive sensitivity.

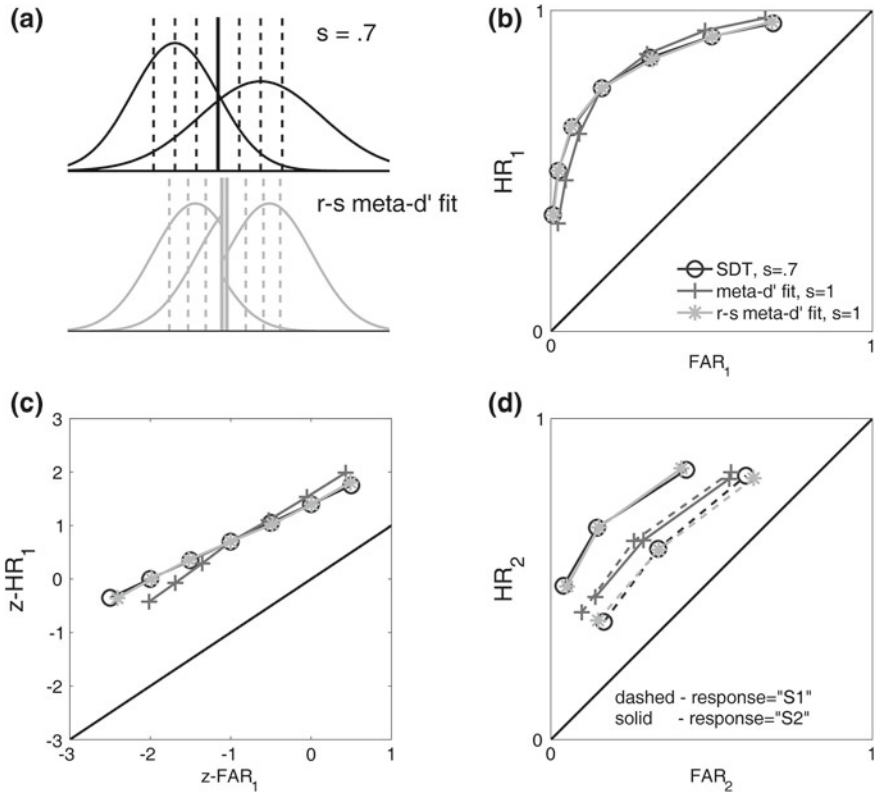


**Fig. 3.6** Response-specific meta-$d'$ model can fit patterns generated by the unequal variance SDT model. **a** UV-SDT model and response-specific meta-$d'$ fit using EV-SDT. We used simulated trials from a UV-SDT model with $s = 0.7$ to generate type 1 and type 2 ROC curves. The response-specific meta-$d'$ fit was able to emulate the differences in the degree of distribution overlap for "*S*1" and "*S*2" responses exhibited by the UV-SDT model (compare distribution overlaps on either side of the type 1 criterion in the top and bottom panels). **b** Type 1 ROC curve. We constructed pseudo type 1 ROC curves from the type 2 (FAR, HR) data produced by the meta-$d'$ fits and the type 1 (FAR, HR) computed from the simulated data according to EV-SDT. Differences between UV-SDT and the meta-$d'$ fits are difficult to discern on the pseudo type 1 ROC. **c** Type 1 z-ROC curve. On the pseudo type 1 $z$-ROC curve it is apparent that UV-SDT produces a curve with a non-unit slope, and that the curve based on response-specific meta-$d'$ under EV-SDT produced a close match. By contrast, the curve based on the meta-$d'$ fit under EV-SDT produced a unit slope. **d** Response-specific type 2 ROC curves. Under the UV-SDT model, there is more area under the type 2 ROC curve for "*S*2" responses than there is for "*S*1" responses. This pattern is closely connected to the non-unit slope on the type 1 $z$-ROC curve. As expected, response-specific meta-$d'$ but not overall meta-$d'$ produced a good fit to this type 2 data

### 3.5.4 RSM-SDT Fit to Data Generated by UV-SDT

We will illustrate the ability of differences in response-specific meta-$d'$ to produce a non-unit slope on the pseudo type 1 z-ROC curve by simulating data from the UV-SDT model and fitting it with RSM-SDT. We used the UV-SDT model with $d'_1 = 2$, $c_1 = 0$, $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.5, -1, -1.5)$, $\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.5, 1, 1.5)$, and $s = 0.7$, where the “1” subscript for $d'$ and $c$ denotes that these are measured in $\sigma_{S1}$ units. The SDT graph for these parameter values is plotted in Fig. 3.6a. We simulated 10,000,000 trials and constructed the pseudo type 1 ROC curve, pseudo type 1 z-ROC curve, and response-specific type 2 ROC curves, as plotted in Fig. 3.6b–d.

Next, we performed both an overall meta-$d'$ fit and a response-specific meta-$d'$ fit to the data, both times using the EV-SDT model as a basis. Performing the meta-$d'$ fit requires first calculating $d'$ and $c$ for the simulated data. Performing the calculations for $d'$ and $c$ under the EV-SDT model yielded $d' = 1.7$ and $c = 0.15$.[15] The overall meta-$d'$ fit resulted in parameter values of meta-$d' = 1.47$, meta-$c = 0.13$, $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.29, -0.74, -1.20)$, and $\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.51, 0.86, 1.20)$. The response-specific meta-$d'$ fit resulted in parameter values of meta-$d'_{\text{“}S1\text{”}} = 1.05$, meta-$c_{\text{“}S1\text{”}} = 0.09$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.28, -0.69, -1.13)$ for “$S1$” responses, and meta-$d'_{\text{“}S2\text{”}} = 2.40$,[16] meta-$c_{\text{“}S2\text{”}} = 0.21$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (0.65, 1.06, 1.45)$ for “$S2$” responses. From these parameter values, we computed the theoretical response-specific type 2 ROC curves (Fig. 3.6d). We also constructed the theoretical pseudo type 1 ROC curves (Fig. 3.6b, c) for the meta-$d'$ fits. It was not possible to do this directly, since the meta-$d'$ fits are meant to describe type 2 performance rather than type 1 outcomes. Thus, we performed the following procedure. From the meta-$d'$ fits, we obtained a set of response-specific ($FAR_2$, $HR_2$) pairs. From the simulated data, we computed the “true” ($FAR_1$, $HR_1$) pair. Then we computed a set of pseudo type 1 ROC points, ($FAR_{1\sim2}$, $HR_{1\sim2}$), using the equations above that describe how to derive pseudo type 1 ROC points from ($FAR_1$, $HR_1$) and a set of response-specific ($FAR_2$, $HR_2$).

Figure 3.6c shows that the UV-SDT model produced a linear z-ROC curve with a slope lower than 1. It also demonstrates that the RSM-SDT fit produced a close approximation to the UV-SDT data, whereas the overall meta-$d'$ fit did not. To quantify these observations, we performed maximum likelihood fits of the UV-SDT model onto (1) the simulated data originally generated by the UV-SDT model, and (2) a new set of 10,000,000 simulated trials that followed a distribution

[15] Note that the values for $d'$ and $c$ recovered by EV-SDT analysis are slightly different from those used in the generating UV-SDT model due to their differing assumptions about the distribution variances.

[16] The value of meta-$d'_{\text{“}S2\text{”}}$ at 2.4 was substantially larger than the value of $d'$ at 1.7, an unusual result as we would typically expect meta-$d' \leq d'$ [13]. However, constraining the RSM-SDT fit such that meta-$d'_{\text{“}S2\text{”}} \leq d'$ still produced data that gave a reasonable approximation to the z-ROC curve. Fitting the UV-SDT model to the data distributed according to this RSM-SDT fit yielded $s = 0.83$, demonstrating that even with the constraint that meta-$d'_{\text{“}S2\text{”}} \leq d'$, RSM-SDT still produced a z-ROC curve with non-unit slope.

of outcomes following the theoretical pseudo type 1 ROC curve generated by the RSM-SDT fit, and (3) similarly for the overall meta-$d'$ fit. The UV-SDT fit to the UV-SDT generated data yielded $s = 0.7$, successfully recovering the true value of $s$ in the generating model. The UV-SDT fit to the data distributed according to RSM-SDT yielded a closely matching $s = 0.72$. The UV-SDT fit to the data distributed according to the overall meta-$d'$ fit yielded $s = 0.98$ since this model has no mechanism with which to produce non-unit slopes on the z-ROC curve.

The relationship between the slope of the pseudo type 1 z-ROC curve and area under the response-specific type 2 ROC curves is made evident in Fig. 3.6d. The data generated by the UV-SDT model produced a type 2 ROC curve for "*S*2" responses that has substantially more area underneath it than does the type 2 ROC curve for "*S*1" responses. Intuitively, this is due to the fact that when $s < 1$, the *S*1 and *S*2 distributions overlap less for "*S*2" responses than they do for "*S*1" responses (see Fig. 3.6a). As expected, the response-specific meta-$d'$ fit is able to accommodate this pattern in the response-specific type 2 ROC curves, whereas the overall meta-$d'$ fit is not.

### 3.5.5 UV-SDT Fit to Data Generated by RSM-SDT

Just as RSM-SDT can closely fit data generated by UV-SDT, here we show that UV-SDT can produce patterns of data similar to those generated by an RSM-SDT model. For this example, we once again use the model described in the section "Toy example of response-specific meta-$d'$ fitting." This model has two parameters, $\sigma_{2,\text{“}S1\text{”}}$ and $\sigma_{2,\text{“}S2\text{”}}$, that control the level of noisiness in type 2 sensitivity for "*S*1" and "*S*2" responses. We simulated 10,000,000 trials using parameter values $d' = 2$, $c = 0$, $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.5, -1, -1.5)$, $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (0.5, 1, 1.5)$, $\sigma_{2,\text{“}S1\text{”}} = 1.5$, and $\sigma_{2,\text{“}S2\text{”}} = 0$.

We fit the RSM-SDT model to this data set, yielding a fit with meta-$d'_{\text{“}S1\text{”}} = 0.78$, meta-$c_{\text{“}S1\text{”}} = 0$, meta-$\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.54, -0.73, -0.94)$ for "*S*1" responses, and meta-$d'_{\text{“}S2\text{”}} = 2.00$, meta-$c_{\text{“}S2\text{”}} = 0$, and meta-$\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.50, 1.00, 1.50)$ for "*S*2" responses. The SDT graphs for these fits are plotted in the top half of Fig. 3.7a.

[17] Note that the nature of the UV-SDT model inherently places constraints upon the set of type 1 and type 2 ROC curves that can be exhibited at the same time, whereas the method for fitting meta-$d'$ minimizes constraints of type 1 performance upon the type 2 fit. Additionally, the likelihood function for the UV-SDT model is built from pseudo type 1 probabilities of the form p(resp = r, conf = y|stim = s). This is different from the likelihood function for fitting meta-$d'$, which is built from type 2 probabilities of the form p(conf = y|stim = s, resp = r). Thus, whereas the meta-$d'$ algorithm is specialized for fitting type 2 data, the fit for the UV-SDT model must account for variance in both type 1 and type 2 responses, entailing potential tradeoffs in the fit. Fitting UV-SDT to the data with a type 2 likelihood function achieves a near perfect fit to the type 2 ROC curves, albeit with a very poor fit to the type 1 ROC curve (data not shown).

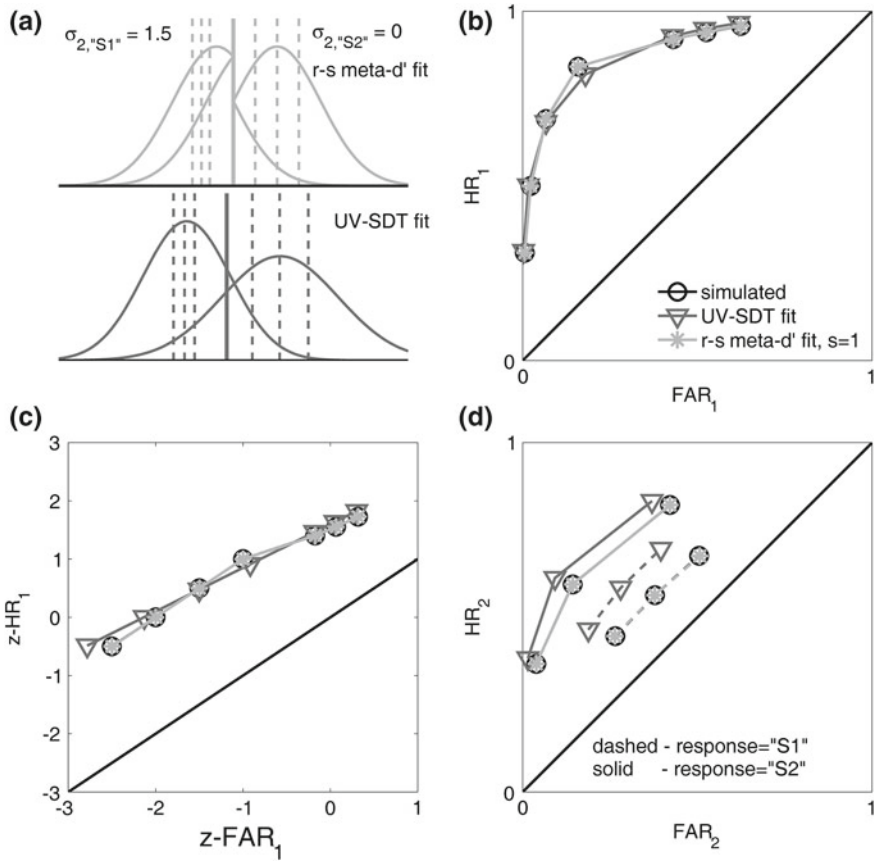


**Fig. 3.7** The unequal variance SDT model can fit patterns generated by asymmetries in response-specific metacognitive sensitivity. **a** Response-specific meta-$d'$ and UV-SDT fits to simulated data. We returned to the model depicted in Fig. 3.5, simulating trials with $\sigma_{2,\text{“}S1\text{”}} = 1.5$ and $\sigma_{2,\text{“}S2\text{”}} = 0$. The *top* half of this panel depicts the response-specific meta-$d'$ fit for the simulated data. The *bottom* half depicts the UV-SDT fit. **b** Type 1 ROC curves. **c** Type 1 z-ROC curves. We produced type 1 ROC curves from the meta-$d'$ fits using the same procedure as in Fig. 3.6. Both the response-specific meta-$d'$ fit and the UV-SDT fit provided a close match to the type 1 ROC curves of the generating model. **d** Response-specific type 2 ROC curves. The UV-SDT model slightly overestimated area under the response-specific type 2 ROC curves, but still captured the fact that there is more area under the curve for “*S*2” responses than for “*S*1” responses

Next, we found the maximum likelihood fit of the UV-SDT model to this data set. This yielded a fit with $d'_1 = 2.14$, $c_1 = -0.15$, $\underline{\boldsymbol{c}}_{2,\text{“}S1\text{”}} = (-0.89, -1.12, -1.37)$, $\underline{\boldsymbol{c}}_{2,\text{“}S2\text{”}} = (0.43, 1.06, 1.72)$, and $s = 0.75$. The SDT graph for this fit is plotted in the bottom half of Fig. 3.7a.

As shown in Fig. 3.7c, the simulated data and meta-$d'$ fit produce a pseudo type 1 z-ROC curve with a small deviation from linearity due to an upward-going kink in the curve corresponding to the “true” (FAR, HR) point. Nonetheless, this curve is closely approximated by the linear z-ROC curve with slope = 0.75 produced by the UV-SDT model fit. The deviation between the UV-SDT fit and the generating model is more clearly pronounced on the response-specific type 2 ROC curves. Although the UV-SDT model overestimates the area under both curves, it nonetheless captures the qualitative pattern that there is more area under the curve for “*S*2” responses than for “*S*1.”[17]

## 3.6 Discussion

### *3.6.1 Implications for Interpretation and Methodology of SDT Analysis of Type 1 and Type 2 Processes*

The foregoing analyses suggest that extra caution should be exercised when interpreting ROC curves. Constructing *z*-ROC curves using confidence rating data risks conflating the contributions of type 1 and type 2 performance. Non-unit slopes on these pseudo type 1 *z*-ROC curves can occur due to response-specific differences in type 2 processing even when the underlying type 1 stimulus distributions have equal variance. Thus, inferences about the nature of type 1 processing based on the pseudo type 1 *z*-ROC curve slope may not always be justified.

This is especially a concern in light of empirical demonstrations that type 1 and type 2 performance can dissociate; e.g., Rounis et al. [17] found that applying transcranial magnetic stimulation to dorsolateral prefrontal cortex selectively diminishes type 2, but not type 1, sensitivity, and Fleming et al. [8] found that between-subject anatomical variation in frontal cortex correlates with variability in type 2 sensitivity even when type 1 sensitivity is held constant across subjects. This suggests that type 2 sensitivity is subject to sources of variation that do not affect type 1 processing. In turn, this suggests that estimates of the relative variances in type 1 stimulus distributions based on the pseudo type 1 ROC curve may be unduly affected by factors that cause variation in type 2, but not type 1, processing.

By the same token, however, differences in response-specific type 2 ROC curves do not necessarily entail differences specifically at the level of type 2 or “metacognitive” processing. Instead, such differences are potentially attributable to differences in basic attributes of type 1 processing, such as type 1 sensitivity, criterion placement, and/or the variability of the type 1 stimulus distributions. For instance, Kanai et al. [11] observed that area under the type 2 ROC curve for

"signal absent" responses were poorer for manipulations that target perceptual, rather than attentional, processes. They inferred that perceptual, but not attentional, manipulations disrupted processing at early levels of processing, such that subjects lacked introspective awareness regarding the source of their failure to detect the target. However, an alternative explanation might be that the type 2 ROC curves differed purely as a consequence of differences in $d'$, $c$, and $s$. Reducing the values of $d'$, $c$, and $s$ can all potentially lead to reductions in area under the type 2 ROC curve for "*S*1" responses. Thus, it is possible that the differences in the type 2 ROC curves for the perceptual and attentional manipulations might be explicable purely in terms of differences in low-level processing, rather than in terms of differences across levels of processing. This is an example of the more general principle upon which our whole approach to type 2 analysis is founded, the principle which necessitates the need for a measure like meta-$d'$: Since type 2 ROC curves depend on the parameters of the type 1 SDT model, it is crucial to interpret type 2 data in the context of the empirical type 1 data, and to consider the extent to which the relationship between the type 1 and type 2 data conforms to or violates SDT expectation [9, 13].

Galvin et al. [9] similarly cautioned against the use of pseudo type 1 ROC curves to make inferences about type 1 processes. They suggested that so-called type 1 ratings (e.g. "rate your confidence that the stimulus was *S*2 on a scale of 1–8") may offer a window into type 1 processing that type 2 ratings (e.g. "rate your confidence that your "*S*1" or "*S*2" response was correct on a scale of 1–4") do not. However, it is not clear that the cognitive mechanisms required to generate such type 1 ratings would differ substantively from those needed for the type 2 ratings, and the informational content of type 1 and type 2 ratings may turn out to be identical, differing only in superficial aspects. In their discussion, Galvin et al. point out that it may be difficult to create a type 2 decision rule that captures the behavior of type 1 ratings. (Per our discussion in the section titled "Comparison of the current approach to that of Galvin et al. [9]", we might say that this is analogous to the problem regarding how to create a type 2 decision rule that adequately captures the empirically observed relationships between the placement of response-specific type 2 criteria.) However, we note that the potential difficulty of such a mapping may simply reflect the possibility that observers do not, in fact, explicitly compute an overall type 2 decision variable as such, or perhaps only do so in a heuristic or variable way.

It may be possible to use z-ROC data to estimate distribution variances without the confounding influence of response-specific type 2 processing by avoiding the use of pseudo type 1 *z*-ROC curves. Instead, type 1 ROC curves can be constructed by using experimental interventions that directly target type 1 decision processes, such as direct instruction, changes in stimulus base rates, or changes in the payoff matrix. On the presumption that such manipulations are not themselves targeting processes that depend on metacognitive or type 2 kind of processing, ROC curves constructed in this way might offer purer windows into the nature of type 1 processing, relatively uncontaminated by the influence of type 2 processing.

This suggestion is consistent with the observation that pseudo type 1 ROC curves do not always behave the same as "true" type 1 ROC curves generated by direct manipulation of the type 1 criterion. For instance, Markowitz and Swets [14] found that estimates of $s$ in auditory detection tasks depend on signal strength for pseudo, but not true, type 1 ROC curves; Van Zandt [23] found that estimates of $s$ based on pseudo type 1 ROC curves varied depending on the degree of bias in the true type 1 criterion (thus implying that not all pseudo type 1 ROC curves yield the same estimate of $s$ as the "true" type 1 ROC curve); and Balakrishnan [1], replicated in Mueller and Weidemann [15], found that pseudo type 1 ROC points can fall below the true type 1 ROC curve constructed under the same experimental conditions. Empirical results like these suggest that pseudo and true-type 1 ROC curves may indeed tap into distinct cognitive processes, which is consistent with our observations that (1) the pseudo type 1 ROC curve has a direct mathematical relationship with response-specific type 2 ROC curves, and (2) type 2 ROC curves are subject to sources of variation that do not affect type 1 performance (e.g. [8, 17]).

These considerations also have implications for the methodology of estimating meta-$d'$. In the current work, and previously, we have considered estimation of meta-$d'$ in the context of equal variance SDT. Only a simple extension of the methodology is needed to perform meta-$d'$ analysis based on the UV-SDT model. Presumably the value of $s$ would be set to a fixed value in the meta-SDT model based on the characteristics of the empirical data being characterized, analogous to the treatment of meta-$c'$. Then the interpretation of meta-$d'$ based upon the UV-SDT model could be expanded to say e.g. "suppose there is an ideal SDT observer O who exhibits a response bias ($c'$) and unequal type 1 variance ($s$) similar to those of subject X. In order for O to produce response-specific type 2 ROC curves like those of X, O would need a $d'$ equal to so-and-so."

However, it is unclear how we could or should arrive at the value of $s$ to be used for such an UV meta-SDT model. As we have seen, the pseudo type 1 ROC curve has a direct mathematical relationship with response-specific type 2 ROC curves, opening up the possibility that measures of $s$ based on the pseudo type 1 ROC curve are confounded by independent sources of variation in type 2 sensitivity. It is not clear that deriving a value for $s$ from pseudo type 1 data, and then using that value of $s$ in a characterization of the type 2 sensitivity exhibited by the very same confidence ratings used to estimate the value of $s$ in the previous step, would be desirable. One potential workaround, as discussed above, might be to independently estimate the type 1 ROC curve based upon direct manipulations of type 1 response bias across experimental conditions. The estimate of $s$ derived from the "true" type 1 ROC curve could then be used to fit an UV meta-SDT model to the type 2 data.

Another option is to gracefully sidestep the problem of UV by utilizing experimental designs that tend to produce data that is adequately characterized by EV-SDT. For example, in 2-interval forced choice designs, the $S1$ stimulus may appear in one of two spatial or temporal intervals, while the $S2$ stimulus appears in the other. The observer must report whether the stimulus sequence on the current

trial was <*S*1, *S*2> or <*S*2, *S*1> (e.g. spatially, "*S*1 was on the left and *S*2 was on the right" or temporally, "*S*2 came first and *S*1 came second"). Intuitively, internal responses should be equally variable for <*S*1, *S*2> and <*S*2, *S*1 > sequences, even if internal responses to *S*1 and *S*2 themselves are not equally variable. This result can be more formally derived from the SDT model [12] and has been observed in empirical data (e.g. [18]). Thus, the 2-inveral forced choice design may provide an experimental paradigm that circumvents concerns related to the UV-SDT model and facilitates usage of EV-SDT.

Another possibility is to create a variation of the SDT model that includes structures to account both for UV and variation in type 2 sensitivity (a simple example of the latter being the $\sigma_2$ model used earlier). It is possible that finding the best fit of such a model to a data set could arbitrate to some extent on the relative contributions of UV and response-specific metacognition to patterns in the data. Such an approach would constitute something of a departure from the meta-$d'$ methodology discussed here. However, it seems likely that such a model-based approach would still need to be supplemented by experimental designs intended to produce data that specifically arbitrate between the mechanisms in question, and it is not clear that the standard form of the two-choice task with confidence ratings provides such a design. Ultimately, analysis of how computational models fit the data needs to be supplemented with other empirical and conceptual considerations in order to make strong inferences about the underlying cognitive processes.

## 3.7 Code for Implementing Overall and Response-Specific Meta-$d'$ Analysis

We provide free Matlab scripts for conducting type 1 and type 2 SDT analysis, including functions to find the maximum likelihood fits of overall and response-specific meta-$d'$ to a data set, at http://www.columbia.edu/∼bsm2105/type2sdt

**Acknowledgements** This work was supported by Templeton Foundation Grant 21569 (H.L.). We thank Dobromir Rahnev and Guillermo Solovey for comments on an earlier version of the manuscript.

## References

1. Balakrishnan JD (1998) Measures and Interpretations of vigilance performance: evidence against the detection criterion. Hum Factors: J Hum Factors Ergon Soc 40(4):601–623. doi:10.1518/001872098779649337
2. Clarke FR, Birdsall TG, Tanner J (1959) Two types of ROC curves and definitions of parameters. J Acoust Soc Am 31(5):629–630. doi:10.1121/1.1907764
3. Clifford CWG, Arabzadeh E, Harris JA (2008) Getting technical about awareness. Trends Cogn Sci 12(2):54–58. doi:10.1016/j.tics.2007.11.009

4. Dorfman DD, Alf E (1969) Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–Rating-method data. J Math Psychol 6(3):487–496. doi:10.1016/0022-2496(69)90019-4
5. Egan JP (1975) Signal detection theory and ROC analysis. Academic Press, New York
6. Evans S, Azzopardi P (2007) Evaluation of a "bias-free" measure of awareness. Spat Vis 20(1–2):61–77
7. Fleming SM, Maniscalco B, Amendi N, Ro T, Lau H (in review). Action-specific disruption of visual metacognition
8. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. Science 329(5998):1541–1543. doi:10.1126/science.1191883
9. Galvin SJ, Podd JV, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychon Bull Rev 10(4):843–876. doi:15000533
10. Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
11. Kanai R, Walsh V, Tseng C-H (2010) Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. Conscious Cogn. doi:10.1016/j.concog.2010.06.003
12. Macmillan NA, Creelman CD (2005) Detection theory: a user's guide, 2nd edn. Lawrence Erlbaum
13. Maniscalco B, Lau H (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn 21(1):422–430. doi:10.1016/j.concog.2011.09.021
14. Markowitz J, Swets JA (1967) Factors affecting the slope of empirical ROC curves: comparison of binary and rating responses. Percept Psychophysics 2(3):91–100. doi:10.3758/BF03210301
15. Mueller ST, Weidemann CT (2008) Decision noise: an explanation for observed violations of signal detection theory. Psychon Bull Rev 15(3):465–494. doi:18567246
16. Ogilvie JC, Creelman CD (1968) Maximum-likelihood estimation of receiver operating characteristic curve parameters. J Math Psychol 5(3):377–391. doi:10.1016/0022-2496(68)90083-7
17. Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci 1(3):165–175. doi:10.1080/17588921003632529
18. Schulman AJ, Mitchell RR (1966) Operating characteristics from Yes-No and Forced-Choice procedures. J Acoust Soc Am 40(2):473–477. doi:10.1121/1.1910098
19. Swets JA (1986) Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull 99(1):100–117. doi:3704032
20. Swets JA (1986) Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 99(2):181–198
21. Swets JA, Tanner WP Jr, Birdsall TG (1961) Decision processes in perception. Psychol Rev 68(5):301–340. doi:10.1037/h0040547
22. Tanner WP Jr, Swets JA (1954) A decision-making theory of visual detection. Psychol Rev 61(6):401–409
23. Van Zandt T (2000) ROC curves and confidence judgements in recognition memory. J Exp Psychol Learn Mem Cogn 26(3):582–600